Iratxe Zenozabal Campillo.
31.10.96.

Mañana: 15'00 - 17'00 Seminario nº 5
Miércoles.

CAPÍTULO III

ESTIMACIÓN DE PARÁMETROS

3.1. Introducción

3.2. Estadístico y estimador

3.3. Principales métodos de estimación

3.4. Estimación por puntos

3.5. Estimación por intervalos

3.6. Estimación por intervalos de la media poblacional

3.7. Estimación por intervalos de la varianza poblacional

TEXTO: Apuntes Fotocopiados

BIBLIOGRAFÍA COMPLEMENTARIA:

BERENSON, M. L. y LEVINE, D. L.: *Estadística para administración y economía*, Madrid, Interamericana, 1987; capítulo 8.

CANAVOS, G. C.: *Probabilidad y estadística*, Madrid, McGraw-Hill, 1988; capítulo 8.

MILLS, R. L.: *Estadística para economía y administración*, Madrid, McGraw-Hill, 1980; capítulo 6.

SPIEGEL, M. R.: *Estadística*, Madrid, McGraw-Hill, 1990; capítulo 9.

ESTIMACIÓN DE PARÁMETROS: EJEMPLOS

EJEMPLO 1:

De una población $N(\mu;\ \sigma = 2)$, donde μ es desconocida, la estimo mediante muestras de tamaño 3. Los estimadores empleados son:

$$A = \frac{X_1 + 2X_2 + X_3}{3}$$

$$B = \frac{X_1 + 2X_2}{3}$$

Comprobar si cumplen las propiedades de los estimadores.

EJEMPLO 2:

Demostrar que la cuasivarianza es un estimador insesgado de σ^2.

EJEMPLO 3:

Dada una variable aleatoria X, que sigue una distribución $N(\mu;\ 3)$, extraigo mediante muestreo aleatorio simple una muestra de tamaño 3, dando como resultado: (12; 14; 16).

Hallar el estimador de μ.

EJEMPLO 4:

De una población normal con $\sigma = 1$ extraigo una muestra de tamaño $n = 50$, obteniendo que $\overline{X} = 10$ y $S = 0,85$.

Hallar un intervalo de confianza del 95% para la media de

la población.

EJEMPLO 5:

Calcular con los datos del **ejemplo 4**, un intervalo de confianza al 95%, considerando que desconozco σ.

EJEMPLO 6:

Se desea estimar la aptitud promedio de determinadas funciones de un considerable grupo de personas, y se quiere que esta estimación se desvíe como mucho en 2 tanto por exceso como por defecto, con una probabilidad de 0,99.

En base a experimentos anteriores con datos promediados, se sabe que σ = 15.

¿Cuál será el tamaño de n?

EJEMPLO 7:

Suponiendo que los pesos netos de las latas de anchoas están distribuidos normalmente, mediante muestreo aleatorio simple tomamos una muestra de n = 12. Los datos obtenidos son:

Grs./lata	439	442	445	448	451	454
Nº latas	1	2	2	3	3	1

Se pide:

a) Estimar el peso medio por lata con un intervalo de confianza del 95%.

b) Estimar σ con un intervalo de confianza del 90%.

EJEMPLO 8:

El salario medio semanal de una fábrica A es de 25.000 pesetas. Tomándose una muestra de n = 31 se obtiene que S_A = 1.400.

En otra fábrica B, el salario medio semanal es de 24.000 pesetas. Tomándose una muestra aleatoria de n = 41 se obtiene que S_B = 1.000.

Suponiendo que los sueldos se distribuyen normalmente, hallar el intervalo de confianza del 99% para estimar la diferencia entre los salarios medios de A y B.

ESTIMACIÓN DE PARÁMETROS: PROBLEMAS

EJERCICIO 1:

Sea X_1, X_2, X_3, X_4, X_5 una muestra aleatoria de una población cuya distribución es normal, con media μ y desviación estándar σ. Considérense los siguientes estadísticos:

$$T1 = \frac{X_1 + X_2 + X_3 + X_4 + X_5}{5}$$

$$T2 = \frac{X_1 + X_2 + X_3 + X_4 + X_5}{6}$$

como estimadores de la media poblacional.

¿Son insesgados? ¿Cuál es más eficiente?

EJERCICIO 2: Véase notas.

Un fabricante de fibras sintéticas desea estimar la tensión media de ruptura de una fibra. Supóngase que la tensión de ruptura de una fibra se encuentra modelada por una distribución normal con desviación estándar conocida de 0,45 libras.

Diseña para ello un experimento en el que se observan las tensiones de ruptura, en libras, de 16 hilos del proceso, seleccionados aleatoriamente. Las tensiones son:

20,8	20,6	21,0	20,9
19,9	20,2	19,8	19,6
20,9	21,1	20,4	20,6
19,7	19,6	20,3	20,7

a) Construir intervalos al nivel de confianza del 90% y 99% para el valor real de la tensión promedio de ruptura de la fibra, comparando los resultados obtenidos.

b) Determinar nuevamente los intervalos de confianza para la tensión media de ruptura de la fibra, sin suponer que se conoce la desviación estándar de la población. ¿Cómo son estos intervalos comparados con los obtenidos anteriormente? ¿Por qué?

t: Dadas los datos. El dueño te los datos y comparar la hipótesis.

EJERCICIO 3:

La Cámara de Comercio de una ciudad se encuentra interesada en estimar la *CANTIDAD TOTAL* que gasta al día la gente (se estima que son unas 5.000 personas) que acude a convenciones, calculando comidas, alojamiento y entretenimiento. De las distintas convenciones que se llevan a cabo en la ciudad, se seleccionaron 16 personas al azar y se les preguntó la cantidad que gastaban al día. Se obtuvo la siguiente información (en u. m.): 150, 175, 163, 148, 142, 189, 135, 174, 168, 152, 158, 184, 134, 146, 155 y 163. Bajo los supuestos que sean necesarios, se pide construir intervalos al 90%, 95% y 99% de nivel de confianza para la cantidad total que gastan al día los 5.000 asistentes a dichas convenciones, comparando los resultados obtenidos.

EJERCICIO 4:

La división de investigación de un departamento estatal de carreteras desea conocer el número promedio de millas

recorridas durante una semana por camiones. El jefe de la división dice que:

- La media de la muestra no deberá alejarse en más de 15 millas, por arriba o por debajo, de la verdadera distancia media recorrida por cada camión.
- El nivel de confianza deberá ser del 95,45%.
- La desviación estándar de la población, basada en estudios previos, se estima que es de 120 millas/semana.

a) ¿Cuál es el tamaño de muestra adecuado para satisfacer estos requerimientos?

b) Suponga que se ha obtenido una muestra aleatoria del tamaño obtenido en el apartado anterior. ¿Considerarías el tamaño de la muestra como satisfactorio si la desviación estándar de la muestra es:

1. 150 millas
2. 96 millas?

Si sí, ¿por qué? Si no, entonces, ¿cuál deberá ser el tamaño de la muestra?

c) El proyecto de investigación presentado en los apartados anteriores fue hecho en un área que tiene 500.000 camiones viajando. Supongamos que las mismas condiciones y requerimientos van a ser aplicados en un área de 2.000.000 camiones. ¿Cómo variará el tamaño de muestra adecuado para el análisis?

EJERCICIO 5:

Una analista desea estimar el monto medio de activos financiados por crédito a largo plazo en una población de 500 empresas. Requiere que la media de la muestra no se aleje, al nivel de confianza del 90%, en más de 50.000$ del auténtico promedio de las 500 empresas. La analista no está muy segura de un valor especificado de la desviación estándar de la población, pero tiene razones para creer que se encuentra entre 100.000$ y 500.000$. La analista tomaría una muestra grande antes que una pequeña. Calcular el correspondiente tamaño de muestra, razonando la respuesta.

EJERCICIO 6:

Supóngase que el funcionario de una línea utiliza la media (27,8 kgs.) de los pesos de una muestra de 50 maletas tomadas al azar entre pasajeros de los vuelos de avión como estimación puntual del peso medio de todas las maletas a transportar hoy en avión.

Si la desviación estándar de la muestra ha sido 5,8 kgs. ¿qué se puede afirmar con probabilidad 0,9 sobre la magnitud máxima del error muestral cometido en la estimación puntual del funcionario?

EJERCICIO 7:

Una muestra aleatoria de 50 botes de piña en rodajas tienen un peso medio de 67,2 onzas y una desviación estándar de 2,1 onzas. ¿Con qué probabilidad se puede decir que la estimación puntual de 67,2 onzas difiere del peso promedio de

todos (50) los botes almacenados en como mucho 0,4 onzas?

EJERCICIO 8:

La calidad de un pedido que está por recibirse depende del número medio de defectos en las piezas del pedido. Se desea obtener un intervalo de confianza para esta media con un error muestral máximo de 0,2 defectos/pieza al nivel de confianza del 95,45%.

De experiencias pasadas se toma el valor de la desviación estándar de la población aproximadamente igual a 1,2 defectos/pieza. Se pide:

a) Determinar el tamaño de muestra a analizar.

b) Una vez analizada la muestra del apartado anterior se obtiene una media muestral de 2 defectos/pieza y una desviación estándar de 1,6 defectos/pieza. Construir el intervalo de confianza para el número medio de defectos/pieza de todo el pedido.

c) El intervalo de confianza calculado, ¿proporciona la precisión deseada? Razónese la respuesta.

d) ¿Qué fue lo que "salió mal" en la planificación previa?

e) Basándote en el intervalo construido, ¿qué conclusiones sacarías respecto a la manera en que se comporta la calidad de este pedido comparada con la de otros pasados, donde se conoce que sólo había por término medio 1 defecto/pieza?

EJERCICIO 9:

El Ayuntamiento quiere estimar el número medio de veces por semana que las casas de la ciudad encienden sus chimeneas. Se sabe que en la ciudad hay alrededor de 20.000 casas con chimenea, por lo que 400 son aleatoriamente entrevistadas con los siguientes resultados: media de encendido 4 veces/semana y desviación estándar 0,6 veces/semana.

Construir un intervalo de confianza al 95% para el número medio de veces que las casas encienden sus chimeneas a la semana.

Si en cada una de las casas se queman exactamente 5 kg. de leña cada vez que se enciende la chimenea, ¿cuál es el intervalo de confianza al 95% para el *CONSUMO TOTAL* de leña a la semana en la citada ciudad?

EJERCICIO 10:

A fin de medir el efecto de una promoción sobre las ventas de una cadena de establecimientos, se seleccionaron al azar 13 de éstos, midiéndose en ellos las ventas antes y después de la campaña promocional.

Los resultados son los que se presentan a continuación.

Tienda	Ventas con campaña	Ventas sin campaña	Tienda	Ventas con campaña	Ventas sin campaña
1	67,2	65,3	8	75,2	69,9
2	59,4	54,7	9	94,7	84,0
3	80,1	81,3	10	64,3	58,4
4	47,6	39,8	11	31,7	33,0
5	97,8	92,5	12	49,3	41,7
6	38,4	37,9	13	54,0	53,6
7	57,3	52,4			

Al nivel de confianza del 95%, ¿podemos decir que la campaña de promoción ha aumentado las ventas medias de la cadena? Razónese la respuesta.

EJERCICIO 11:

Para comparar la efectividad de un programa de seguridad en el trabajo, se observó en seis distintas plantas el número de accidentes por mes antes y después del programa.

Los datos aparecen en la tabla siguiente.

	PLANTA					
	1	2	3	4	5	6
ANTES	38	64	42	70	58	30
DESPUÉS	31	58	43	65	52	29

¿Proporcionan estos datos evidencia suficiente para indicar que el programa ha sido efectivo al reducir el número de accidentes laborales por mes? Trabaje con un nivel de confianza del 90%.

EJERCICIO 12:

La productividad en el trabajo (tiempo necesario para hacer una tarea) depende de muchos y muy variados factores: salario, complejidad de la tarea, ambiente de trabajo... Pero es a menudo el diseño de la propia tarea el factor más importante para la productividad.

Dos diseños de tareas se someten a consideraciones para ser implantados en una fábrica. De un estudio muestral de tiempos y movimientos se obtiene que 12 trabajadores usando el diseño A emplean como media 304 segundos en realizar la tarea, con desviación estándar de 18 segundos. Por su parte, 15 trabajadores usando el diseño B para la misma tarea obtienen una media muestral de 335 segundos, con una desviación estándar de 24 segundos.

Al nivel de confianza del 99%, ¿presentan estos datos evidencia suficiente de igualdad en la tasa media de productividad de ambos diseños?

EJERCICIO 13:

Un nuevo reglamento sobre la emisión máxima de ruido permisible en camiones y autobuses de más de 5 Tn. en tráfico interestatal limita el número de decibelios a 90, medidos a 15 metros, en zonas con velocidad máxima superior a 60 Km/h.

Con objeto de estudiar si durante la instrumentación del nuevo reglamento se tendrá que enfatizar más el control entre el gremio de transportes cargueros o el de compañías de autobuses, se realizó un experimento consistente en la medición, a 15 metros, del ruido emitido por 50 camiones y 50

autobuses en zonas de velocidad máxima superior a 60 Km/h. La media y la desviación estándar muestrales del número de decibelios para los camiones fueron 91,5 y 5,2 decibelios respectivamente, y para los autobuses 88,6 y 3,8. Construir el intervalo de confianza al 95% para la diferencia entre las emisiones medias de ruido de camiones y de autobuses en las circunstancias descritas.

EJERCICIO 14:

En el trabajo de un laboratorio es deseable verificar cuidadosamente la variabilidad de las lecturas obtenidas por un mismo aparato.

En un estudio de la concentración de calcio en agua potable como parte de la valoración de la calidad del agua, se pasó el mismo patrón de medida seis veces en el laboratorio, en intervalos aleatorios. Las lecturas, en partes de millón, fueron: 9,54; 9,61; 9,32; 9,48; 9,70 y 9,26.

Estimar la varianza poblacional para las lecturas obtenidas del patrón, con un nivel de confianza del 90%.

SOLUCIONES

1) a) Sí; b) T_1.

2) a) $P\ [20,1962 \leq \mu_x \leq 20,5664] = 0,9$;

$P\ [20,0916 \leq \mu_x \leq 20,6709] = 0,99$.

b) $P\ [20,1520 \leq \mu_x \leq 20,6105] = 0,9$;

$P\ [19,9959 \leq \mu_x \leq 20,77] = 0,99$.

3) $P\ [151,3057 \leq \mu_x \leq 165,6943] = 0,9$;

$P\ [149,753 \leq \mu_x \leq 167,247] = 0,95$;

$P\ [146,408 \leq \mu_x \leq 170,592] = 0,99$.

4) a) 256; b) Para 150 millas: No; para 96 millas: Sí; n = 400; c) No varía.

5) n = 176.

6) $d_{máx} = 1,3752$.

7) 84,94%.

8) a) n = 144; b) $P\ [1,73 \leq \mu_x \leq 2,27\ ] = 0,9545$;

c) No. 2,6 > 2; d) Varianza infravalorada;

e) Disminución calidad.

9) a) $P\ [4,9412 \leq \mu_x \leq 5,0588] = 0,95$;

b) $P\ [24,706 \leq \mu_y \leq 25,294] = 0,95$.

10) $P\ [1,796 \leq \mu_d \leq 6,2799] = 0,95 \rightarrow$ Sí (intervalo positivo).

11) P $[1{,}3463 \leq \mu_x - \mu_y \leq 6{,}6537] = 0{,}9$ → Sí (intervalo positivo).

12) P $[-54{,}28 \leq \mu_x - \mu_y \leq -7{,}72] = 0{,}99$ → Tiempo B > Tiempo A.

13) P $[1{,}0705 \leq \mu_x - \mu_y \leq 4{,}72948] = 0{,}95$ → Camiones > Autobuses.

14) P $[0{,}0129 \leq \sigma_X^2 \leq 0{,}1247] = 0{,}9$.

TEMA 2

ESTIMACION DE PARAMETROS EN EL MUESTREO ALEATORIO SIMPLE

I Introducción, terminología y conceptos básicos.

II Estadístico y estimador.

III Principales métodos de estimación.

IV Estimación por puntos.

- A. El error muestral.
- B. Propiedades de los estimadores puntuales.
- C. Principales estimadores puntuales.

V Estimación por intervalos.

- A Expresión de un intervalo.
- B Nivel de confianza.

VI Estimación por intervalos de la media poblacional.

- A Situaciones de varianza poblacional conocida.
- B Situaciones de varianza poblacional desconocida.
- C Cálculo del tamaño de muestra.
- D Estimaciones que involucran a dos poblaciones.
  - 1 Muestras independientes.
  - 2 Muestras dependientes.

VII Estimación por intervalos de la proporción poblacional de éxitos.

- A Condiciones de aproximación a la distribución normal.
- B Cálculo del tamaño de muestra.
- C Estimaciones que involucran a dos poblaciones.

VIII Estimación por intervalos de la varianza poblacional.

IX Análisis de costes en el muestreo.

I. INTRODUCCION, TERMINOLOGIA Y CONCEPTOS BASICOS

Nuestro punto de partida va a ser una población de tamaño N, esto es, con N observaciones o mediciones de una variable X, de la cual **deseamos conocer una medida resumen o parámetro poblacional.**

Por ejemplo, dada la población formada por el conjunto de edades de todos los habitantes de Guipúzcoa, podemos estar interesados en calcular el parámetro edad media de dichos habitantes.

Para resolver este problema podemos tomar dos caminos. Por una parte, seleccionar una a una todas las personas que habitan en Guipúzcoa, anotar su edad y calcular de este modo el parámetro edad media. Este proceso de análisis exhaustivo de todas las observaciones de la población recibe el nombre de **censo.**

Como segunda alternativa, podemos extraer de la población formada por el conjunto de habitantes de Guipúzcoa una muestra representativa y aleatoria. Esta muestra estará compuesta por n observaciones, y será sobre ella donde calculemos la edad media. A continuación, y en base al valor del estadístico edad media calculado sobre la muestra, calcularemos o **estimaremos** el valor del parámetro poblacional edad media. A este segundo proceso de calcular el valor del parámetro poblacional desconocido en función del valor conocido de un estadístico en una muestra lo llamamos **estimación.**

Evidentemente, el **censo** es el único modo de llegar a conocer el verdadero valor del parámetro poblacional. En nuestro ejemplo, la única forma de conocer la edad media de todos los habitantes de Guipúzcoa es anotar las edades de todas y cada una de las personas que habitan en Guipúzcoa y calcular la media de dichas edades. Sin embargo, es también claro que la realización de un censo implica unos elevados costes que en ocasiones no se pueden afrontar. Por tanto, rechazamos como solución los censos por razones de coste y de tiempo. También rechazamos los censos por motivos de calidad de información. La realización de un censo implica la utilización de un gran volumen de datos. Esto favorece que se cometan errores. La utilización de muestras implica un menor volumen de datos y por tanto una menor probabilidad de error en su manejo.

La estimación tiene una gran utilidad práctica y salvo algunas excepciones (como es el caso del censo electoral para las votaciones) es el método empleado en todo tipo de estudios de opinión y de mercado.

II. ESTADISTICO Y ESTIMADOR

Como ya quedó definido en el tema 1, un **estadístico** es una medida resumen de la información obtenida en la muestra.

De todos los posibles estadísticos que se pueden calcular en una muestra, **estimador** es el estadístico que empleamos para llevar a cabo la inferencia sobre el valor desconocido del parámetro poblacional.

Por ejemplo, la media muestral, la mediana muestral y la moda muestral son estadísticos, pero sólo utilizaremos la media muestral como estimador de la media poblacional desconocida.

Esto nos lleva a afirmar que **todo estimador es un estadístico** pero **no todo estadístico hará el papel de estimador.**

III. PRINCIPALES METODOS DE ESTIMACION

Existen básicamente dos métodos de estimación, que son la **estimación por puntos** y la **estimación por intervalos.**

La estimación por puntos consiste en dar para el parámetro desconocido que deseamos estimar un único valor, coincidente con el que toma el estimador de dicho parámetro en la muestra.

Siguiendo con el ejemplo de la población de Guipúzcoa, supongamos que una vez extraída una muestra de 250 personas hemos encontrado que el valor del estadístico media muestral es de 50 años. La estimación por puntos consistiría en decir: "Considero que la edad media de **TODOS** los habitantes de Guipúzcoa es de 50 años".

La estimación por intervalos consiste en dar para el parámetro desconocido un conjunto o intervalo de valores. Este conjunto de valores está construido en base al valor que el estimador del parámetro toma en la muestra.

Por ejemplo, volviendo al caso anterior, una estimación por intervalo de la edad media de todos los habitantes de Guipúzcoa podría llevarse a cabo del siguiente modo: "Ya que la media muestral es de 50 años, considero que la edad media de **TODOS** los habitantes de Guipúzcoa se encuentra entre 45 y 55 años".

IV. ESTIMACION POR PUNTOS

A. El error muestral

El error muestral, que denotaremos por d, se define como el error que se comete por el hecho de utilizar muestras para estimar parámetros. El error muestral **es la diferencia, en valor absoluto, entre el valor del estimador y el verdadero valor del parámetro.**

$$d = \text{error muestral}$$

$$d = \left| \begin{array}{c} \text{Valor} \\ \text{del} \\ \text{estimador} \end{array} - \begin{array}{c} \text{valor} \\ \text{del} \\ \text{parámetro} \end{array} \right|$$

Sin embargo, nos encontramos con un problema: si estamos tratando de estimar el valor del parámetro es porque no lo conocemos, y si no conocemos el valor del parámetro **no podremos conocer nunca cuánto vale el error muestral que estamos cometiendo en una estimación puntual.**

Haciendo un símil, podemos decir que la estimación puntual es como lanzar un dardo a una diana (tratar de estimar el parámetro) sin saber dónde está el centro de la diana (sin conocer cuál es el verdadero valor del parámetro). En estas condiciones no podremos saber si un lanzamiento ha acertado o no en el centro de la diana, y si ha errado, en qué magnitud.

Afortunadamente, aunque el error muestral "d" no puede ser calculado, **sí puede ser controlado.** A medida que aumentemos el tamaño de la muestra, la probabilidad de que el estadístico muestral que empleamos como estimador del parámetro poblacional tome valores muy lejanos a este último ira disminuyendo. Dicho en otras palabras, a medida que aumentemos el tamaño de la muestra esperamos que el error muestral disminuya.

Evidentemente, lo ideal es que el error muestral valga 0, ya que queremos que la estimación del valor del parámetro sea lo más exacta posible. El problema es que nunca sabremos si el error muestral es 0, porque siempre desconoceremos cuánto vale d.

B. Propiedades de los estimadores puntuales

De todos los estadísticos que se pueden calcular en la muestra, ¿cuál escogeremos para realizar la estimación puntual del parámetro? Dicho en otras palabras, ¿cuáles son las propiedades que debe reunir un estadístico para que pueda ser considerado un buen estimador puntual de un parámetro poblacional desconocido?

Un buen estimador debe cumplir cuatro requisitos. Debe ser:

* INSESGADO.
* CONSISTENTE.
* EFICIENTE.
* SUFICIENTE.

INSESGADO: Decimos de un estimador que es insesgado **cuando el valor esperado de su distribución muestral coincide con el valor del parámetro que trata de estimar.**

$$E(estimador) = Parámetro$$

Podemos por tanto decir que la media muestral es un estimador insesgado de la media poblacional ya que el valor esperado de la distribución muestral de medias coincide (tal como se comprobó en el tema 1) con el valor de la media poblacional. Podemos también decir que la proporción muestral de éxitos es un estimador insesgado de la proporción poblacional de éxitos por el mismo motivo. La varianza muestral tal como la hemos definido (dividida entre n-1) es un estimador insesgado de la varianza poblacional.

CONSISTENTE: Decimos que un estimador es consistente **cuando la probabilidad de cometer un error muestral determinado disminuye a medida que aumenta el tamaño de la muestra.**

$$\lim_{n\to\infty} P(d > k) \to 0$$

Para que un estimador sea consistente debe ser insesgado y la varianza de su distribución muestral debe tender a 0 a medida que aumenta el tamaño de la muestra.

Podemos afirmar que la media muestral es un estimador consistente del parámetro media poblacional. La proporción muestral de éxitos es un estimador consistente de la proporción poblacional de éxitos. La varianza muestral es un estimador consistente de la varianza poblacional.

EFICIENTE: La eficiencia es un criterio relativo que se emplea para comparar posibles estimadores de un mismo parámetro. **De entre varios estimadores será más eficiente aquel que tenga una distribución muestral con menor varianza.**

SUFICIENTE: Un estimador es suficiente **cuando recoge toda la información de la muestra y es sensible a cambios en la misma.**

Ejemplo: Sean X_1, X_2, X_3 y X_4 cuatro observaciones de una muestra aleatoria simple extraída de una población con media μ y varianza σ^2. Considérense los tres siguientes posibles estimadores de μ:

$$A=\frac{X_1+X_4}{2} \qquad B=\frac{X_1+2X_4}{3}$$

$$C=\frac{X_1+X_2+X_3}{3}$$

¿Son insesgados? ¿Cuál es el más eficiente? ¿Son suficientes?

Para comprobar si son insesgados deberemos calcular su valor esperado:

$$E(A)=E\left(\frac{X_1+X_4}{2}\right)=\frac{1}{2}E(X_1+X_4)=\frac{1}{2}(\mu+\mu)=\mu$$

$$E(B)=E\left(\frac{X_1+2X_4}{3}\right)=\frac{1}{3}E(X_1+2X_4)=\frac{1}{3}(\mu+2\mu)=\mu$$

$$E(C)=E\left(\frac{X_1+X_2+X_3}{3}\right)=\frac{1}{3}E(X_1+X_2+X_3)=\frac{1}{3}(\mu+\mu+\mu)=\mu$$

Dado que el valor esperado de la distribución muestral del estimador coincide con el valor del parámetro que estamos tratando de estimar podemos decir que los tres estimadores son insesgados.

Para conocer cuál de los estimadores es el más eficiente debemos calcular su varianza. Aquel de los estimadores que tenga una menor varianza será el más eficiente de los tres:

$$Var(A) = Var\left(\frac{X_1+X_4}{2}\right) = \frac{1}{4} Var(X_1+X_4) =$$

$$= \frac{1}{4}(\sigma^2+\sigma^2) = \frac{\sigma^2}{2}$$

$$Var(B) = Var\left(\frac{X_1+2X_4}{3}\right) = \frac{1}{9} Var(X_1+2X_4) =$$

$$= \frac{1}{9}(\sigma^2+4\sigma^2) = \frac{5}{9}\sigma^2$$

$$Var(C) = Var\left(\frac{X_1+X_2+X_3}{3}\right) = \frac{1}{9} Var(X_1+X_2+X_3) =$$

$$= \frac{1}{9}(\sigma^2+\sigma^2+\sigma^2) = \frac{\sigma^2}{3}$$

El estimador que presenta una menor varianza es C. Por tanto, diremos que el estimador más eficiente de los 3 es C.

Sin embargo, ninguno de los tres estimadores propuestos es suficiente ya que ninguno de ellos recoge toda la información de la muestra.

Los estimadores primero y segundo dejan de lado la segunda y la tercera observación, por lo que algunos cambios en la muestra (cambios que se produjeran en la segunda y tercera observación) no afectarían al valor del estimador.

Por otra parte, el tercer estimador obvia el valor de la cuarta observación.

C. Principales estimadores puntuales

$$\rightarrow PARAMETRO = \mu_x$$

$$ESTIMADOR = \bar{x}$$

$$E(estimador) = \mu_x$$

$$Var(estimador) = \frac{\sigma^2_x}{n}$$

$$\rightarrow PARAMETRO = P$$

$$ESTIMADOR = \bar{p}$$

$$E(estimador) = P$$

$$Var(estimador) = \frac{PQ}{n}$$

$$PARAMETRO = \sigma^2_x$$

$$ESTIMADOR = s^2$$

$$E(estimador) = \sigma^2_x$$

$$Var(estimador) = \frac{2\sigma^4_x}{n-1}$$

$$PARAMETRO = \mu_x - \mu_y$$

$$ESTIMADOR = \bar{x} - \bar{y}$$

$$E(estimador) = \mu_x - \mu_y$$

$$Var(estimador) = \frac{\sigma^2_x}{n_x} + \frac{\sigma^2_y}{n_y}$$

$$PARAMETRO = P_x - P_y$$

$$ESTIMADOR = \bar{p}_x - \bar{p}_y$$

$$E(estimador) = P_x - P_y$$

$$Var(estimador) = \frac{P_xQ_x}{n_x} + \frac{P_yQ_y}{n_y}$$

V. ESTIMACION POR INTERVALOS

A. Expresión de un intervalo (=centro ± radio).

Recordemos que la estimación por intervalo consiste en dar un conjunto o intervalo de valores como estimación del valor del parámetro poblacional desconocido.

La mayor parte de los intervalos empleados en este tipo de estimación están formados por un **centro** y un **radio**, que se resta y suma al centro del intervalo para obtener el límite inferior y el límite superior del intervalo.

El centro del intervalo coincide con el **valor que en la muestra ha tomado el estimador puntual** del parámetro, mientras que el radio expresa el **ERROR MUESTRAL MAXIMO QUE SE DESEA COMETER** (esto es, la máxima diferencia en valor absoluto que se desea entre el valor del estimador puntual y el auténtico valor del parámetro desconocido).

El radio de un intervalo puede también expresarse como el **producto entre una constante y la desviación estándar del estimador puntual.** Por ejemplo, el intervalo que emplearemos para estimar el valor del parámetro media poblacional tendrá la siguiente forma:

$$\overline{X} \pm kte\sqrt{\frac{\sigma^2_x}{n}}$$

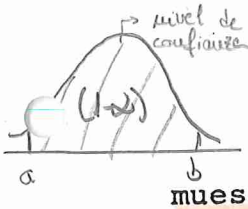
Y el intervalo que emplearíamos para estimar el valor del parámetro proporción poblacional de éxitos se expresaría así:

$$\overline{p} \pm kte\sqrt{\frac{PQ}{n}}$$

Una propiedad deseable de los intervalos que construyamos es que tengan un radio lo más reducido posible. Intervalos con radios muy amplios hacen que la estimación resulte muy ambigua y, por tanto, de escasa utilidad. Por ejemplo, preferiremos que alguien

nos diga "La edad media de los habitantes de Guipúzcoa se encuentra entre 48 y 52 años" a que nos diga "La edad media de los habitantes de Guipúzcoa se encuentra entre 25 y 75 años" (esto último ya lo podíamos sospechar y nos aporta mucha menos información que la primera afirmación).

Dicho en otras palabras, **deseamos que nuestras estimaciones cometan un error muestral máximo lo más reducido posible.**



B. Nivel de confianza (=1-α)

probabilidad de q. el parámetro poblacional quede incluido en ese intervalo.

Por el hecho de no hacer censos y estar trabajando con muestras, **jamás tendremos la certeza absoluta de que el intervalo construido en base a un estadístico muestral incluya entre sus límites el verdadero valor desconocido del parámetro poblacional.** Si no tenemos certeza existe incertidumbre, que se manifiesta en términos de probabilidad.

A la probabilidad de que el intervalo construido en base al estimador puntual incluya al parámetro poblacional desconocido se le llama NIVEL DE CONFIANZA.

El nivel de confianza será técnicamente designado como $1-\alpha$. Lógicamente, desearemos que la probabilidad de que el intervalo construído encierre entre sus límites al parámetro poblacional **sea lo más alta posible**, es decir, desearemos que el nivel de confianza sea muy próximo a 1 (aunque jamás podrá llegar a alcanzar ese valor).

↳ error muestral mínimo.

Los niveles de confianza empleados habitualmente suelen ser 0,90; 0,95; 0,9545; 0,99 y 0,9997.

Sin embargo, los dos objetivos que hemos manifestado como deseables en todo intervalo (a saber, un error muestral máximo lo más reducido posible y un nivel de confianza muy próximo a 1) **entran en contradicción.**

Si ampliamos el radio del intervalo (aumentamos el error muestral máximo) tendremos un mayor nivel de confianza (si aumentamos el radio del intervalo éste encerrará ahora más valores, y la probabilidad de que se incluya al parámetro poblacional desconocido será mayor).

Si reducimos el radio del intervalo (reducimos el error muestral máximo), estaremos automáticamente reduciendo el nivel de confianza (el intervalo tendrá menos valores, y por tanto será ahora menos probable haber encerrado entre sus límites al parámetro poblacional).

La única forma de satisfacer ambos objetivos simultáneamente es incrementar el tamaño de muestra a analizar. Esto es, para un nivel de confianza dado, podremos aumentar o disminuir el error muestral máximo mediante un decremento o un incremento respectivamente en el tamaño de la muestra. Y para un error muestral máximo dado, podremos aumentar o disminuir el nivel de confianza aumentando o disminuyendo el tamaño de la muestra.

VI. ESTIMACION POR INTERVALOS DE LA MEDIA POBLACIONAL

A. Situaciones de varianza poblacional conocida

Nuestro punto de partida es una población representada por N mediciones de una variable aleatoria X, de la que conocemos que su naturaleza es normal y **el valor de su varianza.** Por el contrario, desconocemos el valor del parámetro media poblacional (μ) y deseamos estimarla por intervalos.

En casos como este, sabemos que el comportamiento de la distribución muestral de medias es el siguiente:

Partimos de $X \Rightarrow N(\mu_x ; \sigma^2_x)$

donde $\mu_x = ?$

pero σ^2_x = *Valor conocido*

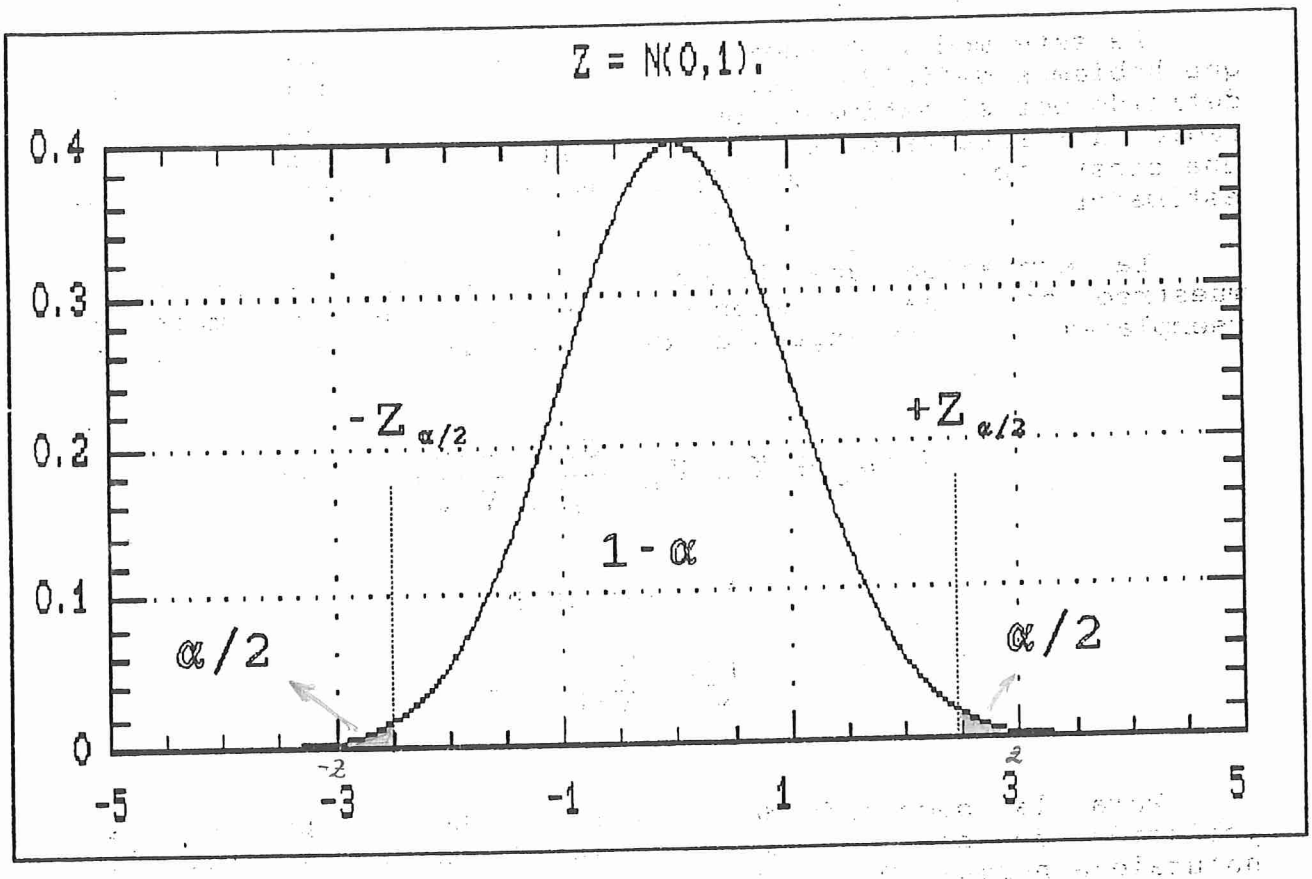
Sabemos que: $\overline{X} \Rightarrow N(\mu_x ; \frac{\sigma^2_x}{n}) \Rightarrow$

Primero tipificaremos pa' trabajar con puntuaciones "Z".

$$\Rightarrow \frac{\overline{X} - \mu_x}{\frac{\sigma_x}{\sqrt{n}}} = Z$$

Pretendemos establecer un intervalo de confianza.

Dado un nivel de confianza $1-\alpha$, se cumplirá que:



$$P\ (-Z_{\alpha/2} \leq Z \leq +Z_{\alpha/2}) = 1-\alpha$$

$$P\ \left(-Z_{\alpha/2} \leq \frac{\overline{X}-\mu_x}{\frac{\sigma_x}{\sqrt{n}}} \leq +Z_{\alpha/2}\right) = 1-\alpha$$

Para pasar a operaciones normales destipifico

Operando para aislar μ_x:

$$P\ \left(\overline{X} - Z_{\alpha/2}\frac{\sigma_x}{\sqrt{n}} \leq \mu_x \leq \overline{X} + Z_{\alpha/2}\frac{\sigma_x}{\sqrt{n}}\right) = 1-\alpha$$

O lo que es igual:

$$P\ \left(\mu_x \in \overline{X} \pm Z_{\alpha/2}\frac{\sigma_x}{\sqrt{n}}\right) = 1-\alpha$$

radio o error muestral. $d = \left| Z_{\alpha/2}\frac{\sqrt{x}}{\sqrt{n}} \right|$

Si $\sigma_x \uparrow$, $d \uparrow$
Si aumenta el tamaño de la muestra, disminuye d:
Si aumenta $(1-\alpha)$, aumentará $d: \left| Z_{\alpha/2} \frac{\sigma_x}{\sqrt{n}} \right|$

De este modo, obtenemos un intervalo que responde a la forma que habíamos definido previamente. El centro del intervalo viene definido por el estimador puntual, media muestral, mientras que el radio del intervalo (el error muestral máximo que se comete) es una constante ($Z_{\alpha/2}$) multiplicada por la desviación estándar del estimador.

La expresión que hemos hallado es válida sólamente para muestreo con reemplazamiento. En el caso de muestreo sin reemplazamiento, la expresión del intervalo varía ligeramente:

$$P\left(\mu_x \in \bar{X} \pm Z_{\alpha/2} \frac{\sigma_x}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}\right) = 1-\alpha$$

Nótese que:

$$\lim_{N \to \infty} \frac{N-n}{N-1} = 1$$

Para la deducción del intervalo hemos supuesto que en la población de partida la variable X objeto de estudio tenía naturaleza normal ¿Qué ocurre si esto no es así? En estos casos la expresión del intervalo deducida es válida **siempre y cuando se cumpla el teorema central del límite.**

Una pregunta interesante que podemos plantearnos es la siguiente: ¿de qué depende la longitud del radio del intervalo? Dicho en otras palabras, ¿de qué depende el error muestral máximo que estamos cometiendo en una estimación del parámetro media poblacional? La respuesta la tenemos en

$$d = \text{ERROR MUESTRAL MAXIMO}$$

$$d = \left| \begin{array}{c} \text{Valor} \\ \text{del} \\ \text{estimador} \end{array} - \begin{array}{c} \text{valor} \\ \text{del} \\ \text{parámetro} \end{array} \right|$$

$$d = \left| Z_{\alpha/2} \frac{\sigma_x}{\sqrt{n}} \right|$$

Podemos por tanto decir que el error muestral que se comete en una estimación depende:

- Del nivel de confianza utilizado, en una relación directa.
- De la dispersión o variabilidad de la variable analizada en la población, en una relación directa.
- Del tamaño de muestra analizado, en una relación inversa.

Por último, es importante señalar que en los casos de varianza poblacional conocida, dados **un nivel de confianza** y **un tamaño de muestra** determinados, todos los intervalos que podamos calcular tratando de estimar el verdadero valor del parámetro media poblacional tendrán **centro variable** (el centro depende del valor que el estadístico media muestral tome en la muestra) pero **radio constante** (ya que el radio no depende de ningún estadístico calculado en la muestra).

Ejemplo: Las horas/mes de absentismo entre el personal de una gran empresa siguen una distribución normal, con desviación estándar conocida de 3 horas/mes y media desconocida. Analizada una muestra aleatoria de 100 trabajadores se obtuvo una media muestral de 10 horas/mes de absentismo.

a) Realizar una estimación puntual de la media poblacional de horas/mes de absentismo para todos los trabajadores de la empresa.

b) Hacer una estimación por intervalos al 90% y al 99% de la media poblacional de horas/mes de absentismo de todos los trabajadores de la empresa.

X: Horas/mes de absentismo de 1 trabajador de la empresa.

Nat(X) --> NORMAL μ_x = ? σ_x = 3 horas/mes

De la población formada por el conjunto de trabajadores de la empresa (que podemos considerar a todos los efectos infinita), hemos tomado una muestra de tamaño n=100 mediante muestreo sin reemplazamiento. Dado el tamaño de la población, dispondremos de un conjunto de 100 VAIID (tal y como se vio en el tema anterior).

Sobre esta muestra calculamos la media muestral y obtenemos que, para la muestra concreta que hemos elegido, la media muestral toma un valor de 10 horas/mes. De este modo, la estimación puntual de la media poblacional del número de horas/mes de absentismo laboral dirá **"El promedio de horas/mes de absentismo laboral para TODOS los trabajadores de la empresa es 10".**

$$\hat{\mu}_x = \overline{X} = 10 \ horas/mes$$

Para obtener los intervalos que se solicitan emplearemos la expresión que hemos deducido previamente:

$$\overline{X} \pm Z_{\alpha/2}\frac{\sigma_x}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} \approx \overline{X} \pm Z_{\alpha/2}\frac{\sigma_x}{\sqrt{n}}$$

$$Para\ 1-\alpha = 0,90$$

$$10 \pm 1,645\frac{3}{\sqrt{100}} = 10 \pm 0,49 \Rightarrow$$

$$\Rightarrow P\ (9,51 \leq \mu_x \leq 10,49) = 0,90$$

Una vez que hemos obtenido este intervalo no sabemos con certeza si hemos acertado o hemos fallado en nuestra predicción. Es decir, **no podemos afirmar con absoluta certeza que la verdadera media poblacional se encuentre entre 9,51 y 10,49.**

Lo que sí podemos afirmar con una probabilidad del 90% es que el error muestral máximo, esto es, la máxima diferencia entre el verdadero valor del parámetro μ_x y el valor de la media muestral es 0,49 horas/mes.

Para un nivel de confianza del 99% tendremos:

$$\overline{X} \pm Z_{\alpha/2}\frac{\sigma_x}{\sqrt{n}} = 10 \pm 2,575\frac{3}{\sqrt{100}} =$$

$$= 10 \pm 0,7725$$

$$P\ (9,2275 \leq \mu_x \leq 10,7725) = 0,99$$

Al igual que en el caso anterior, no podemos afirmar con absoluta certeza que la media poblacional esté comprendida en el intervalo que hemos calculado. Podemos sin embargo afirmar que con una probabilidad del 99% la media poblacional no se alejará de la media muestral calculada en más de 0,7725 horas/mes.

Como podemos comprobar, si deseamos tener una mayor seguridad de haber capturado en el intervalo a la media poblacional (esto es, si deseamos un mayor nivel de confianza) debemos ampliar la amplitud del intervalo. Para obtener certeza absoluta deberíamos ampliar el intervalo desde menos infinito hasta más infinito.

B. Situaciones de varianza poblacional desconocida

Partimos del estudio de una población formada por N mediciones de una variable aleatoria X, que sigue una distribución normal y de la cual desconocemos tanto su media como su varianza.

En casos como este, sabemos que el comportamiento de la distribución muestral de medias es el siguiente:

$$\text{Partimos de} \quad X \Rightarrow N(\mu_x ; \sigma^2{}_x)$$

$$\text{donde } \mu_x = ?$$

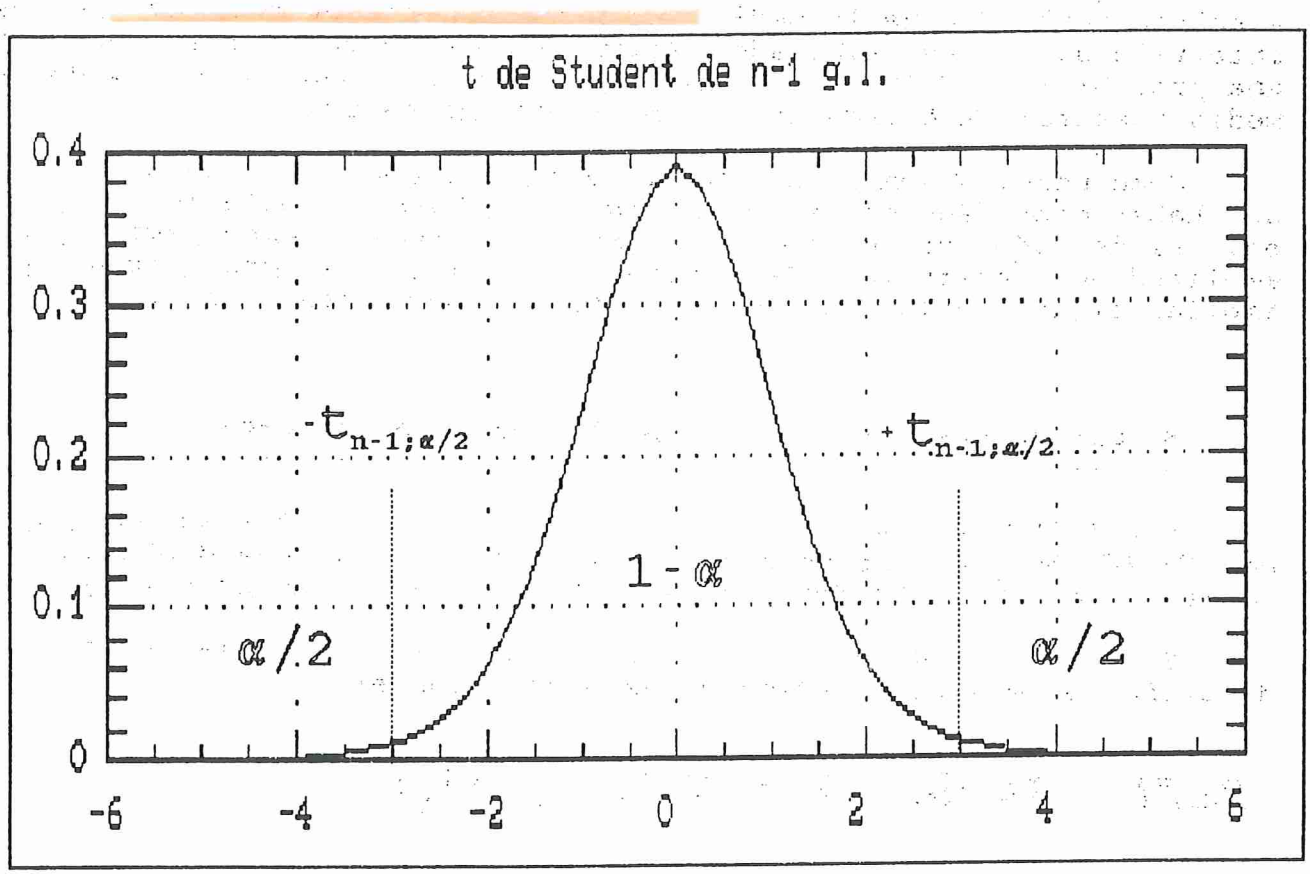
$$\text{y } \sigma^2{}_x = ?$$

$$\text{Sabemos que: } \bar{X} \Rightarrow N\left(\mu_x ; \frac{\sigma^2{}_x}{n}\right) \Rightarrow \hat{\sigma}_x{}^2 = s^2$$

$$\text{Por tanto: } \bar{X} \Rightarrow N\left(\mu_x ; \frac{s^2}{n}\right) \Rightarrow$$

$$\Rightarrow \frac{\bar{X} - \mu_x}{\frac{s}{\sqrt{n}}} = t_{n-1}$$

Dado un nivel de confianza 1-α, se cumplirá que:



$$P\ (-t_{n-1;\alpha/2} \leq t_{n-1} \leq +t_{n-1;\alpha/2}) = 1-\alpha \Rightarrow$$

$$P\ (-t_{n-1;\alpha/2} \leq \frac{\overline{X}-\mu_x}{\frac{S}{\sqrt{n}}} \leq +t_{n-1;\alpha/2}) = 1-\alpha \Rightarrow$$

$$P(\overline{X} - t_{n-1;\alpha/2}\frac{S}{\sqrt{n}} \leq \mu_x \leq \overline{X} + t_{n-1;\alpha/2}\frac{S}{\sqrt{n}}) = 1-\alpha$$

O lo que es igual:

$$P\ (\mu_x \in \overline{X} \pm t_{n-1;\alpha/2}\frac{S}{\sqrt{n}}) = 1-\alpha$$

Como vemos, el intervalo definido responde una vez más a la forma de (centro ± radio). El centro del intervalo es el estimador puntual (media muestral) del parámetro poblacional (media poblacional) que deseamos estimar. El radio del intervalo (el máximo error muestral que se comete en la estimación) tiene la forma de una constante (una t Student de n-1 grados de libertad) multiplicada por **la estimación del error estándar del estimador**, ya que el verdadero valor del error estándar del estimador es desconocido, al ser desconocido el valor de la varianza poblacional.

El intervalo definido es válido para los casos de muestreo con reemplazamiento. En el caso de **muestreo sin reemplazamiento**, la forma del intervalo varía ligeramente:

$$P(\mu_x \in \overline{X} \pm t_{n-1;\alpha/2}\frac{S}{\sqrt{n}}\sqrt{1-\frac{n}{N}}) = 1-\alpha$$

Nótese que:

$$\lim_{N\to\infty}(1-\frac{n}{N}) = 1$$

Recordemos además que, en el caso de que el tamaño de muestra analizado para construir el intervalo sea muy grande, **la t Student con muchos grados de libertad tiende hacia la normal Z.** Es por ello que para tamaños de muestra grandes, los intervalos definidos anteriormente se pueden aproximar a:

$$P(\mu_x \in \overline{X} \pm Z_{\alpha/2}\frac{S}{\sqrt{n}}) = 1-\alpha \quad MCR$$

$$P(\mu_x \in \overline{X} \pm Z_{\alpha/2}\frac{S}{\sqrt{n}}\sqrt{1-\frac{n}{N}}) = 1-\alpha \quad MSR$$

En el caso de que la naturaleza de la variable que se estudia en la población de partida **no fuese normal**, el intervalo anterior es válido si el tamaño de muestra es muy grande. En la práctica son necesarias al menos de 50 a 100 observaciones, según lo simétrica o asimétrica que sea la distribución de la variable de partida en la población.

Por último, es necesario señalar que en el caso de varianza poblacional desconocida, **dados un nivel de confianza** y **un tamaño de muestra** determinados, los intervalos que podamos construir tendrán además de **centro variable** (ya que el estadístico media muestral puede variar de muestra a muestra) un **radio variable.**

En el caso de varianza poblacional desconocida, el radio del intervalo viene determinado por **una estimación del error estándar del estimador.** Dado que esa estimación se lleva a cabo sobre la muestra extraída, podrá variar también de muestra a muestra. Esto supone que el radio del intervalo pueda variar de una muestra a otra.

Ejemplo: Un director de personal desea estimar la calificación media obtenida en un test de selección de candidatos a un puesto de trabajo al que se han presentado 200.000 personas. Para ello selecciona al azar una muestra de 1210 exámenes, obteniendo en ellos una media muestral de 75,6 puntos sobre 100 y una varianza muestral de 216 puntos². Se pide realizar una estimación puntual de la calificación media obtenida, así como el intervalo de estimación para un nivel de confianza del 95%

X: Puntuación obtenida en el test por 1 candidato

Nat X = ? μ_x=? σ_x =?

Tomamos de la población una muestra aleatoria de tamaño 1210. Disponemos de este modo de 1210 VAIID. Sobre esta muestra calculamos los estadísticos media muestral y varianza muestral y encontramos que, para la muestra concreta que hemos extraído, sus valores son 75,6 puntos y 216 puntos² respectivamente.

La estimación puntual del parámetro media poblacional vendrá dada por el valor que en la muestra ha tomado el estimador de la media poblacional, esto es, la media muestral:

$$\hat{\mu}_x = \overline{X} = 75,6 \; puntos$$

De este modo afirmamos: "La calificación media obtenida por **los 200.000 candidatos** en el test de selección ha sido de 75,6 puntos".

Para hallar el intervalo al nivel de confianza del 95% emplearemos la expresión que hemos deducido con anterioridad:

$$P\ (\mu_x \in \bar{X} \pm t_{n-1;\alpha/2}\frac{S}{\sqrt{n}}\sqrt{1-\frac{n}{N}}) = 1-\alpha \Rightarrow$$

¿Cómo la hallado "S"?

$$P(\mu_x \in 75,6 \pm 1,96\frac{14,69}{\sqrt{1210}}\sqrt{1-\frac{1210}{200000}}) = 0,95 \Rightarrow$$

$$P\ (74,78 \leq \mu_x \leq 76,42) = 0,95$$

No podemos asegurar con absoluta certeza que la media poblacional se encuentre dentro del intervalo mencionado. Lo que sí podemos afirmar con una probabilidad de equivocarnos del 5% es que, analizando muestras de 1210 test de un conjunto de 200.000, la diferencia entre la media muestral (estimador de la media poblacional) y la verdadera media poblacional será a lo sumo de 0,82 puntos.

C. Cálculo del tamaño de muestra

Hasta el momento el tamaño de la muestra ha sido siempre un dato en las situaciones analizadas (nos venía dado). Vamos ahora a plantearnos calcular el tamaño de muestra antes de construir el intervalo de confianza.

Hemos visto con anterioridad que los objetivos **alto nivel de confianza** y **reducido error muestral máximo** eran opuestos. También se ha señalado que la forma de satisfacer simultáneamente ambos requerimientos era incrementar el tamaño de la muestra.

Cabe entonces preguntarse: ¿Cuál es el tamaño de muestra necesario para que, dado un nivel de confianza $1-\alpha$, el error muestral en que incurrimos no vaya más allá de $d_{máx}$?

Para determinar el tamaño de muestra es por tanto necesario fijar **subjetivamente** el **nivel de confianza** $(1-\alpha)$ deseado para el intervalo, así como el **error muestral máximo** $d_{máx}$ que estamos dispuestos a cometer.

Debemos distinguir entre muestreo con reemplazamiento y muestreo sin reemplazamiento a la hora de calcular el tamaño de muestra.

Para **muestreo con reemplazamiento** el intervalo de estimación del parámetro media poblacional cuando se conoce la varianza de la población viene dado por la siguiente expresión:

$$\overline{X} \pm Z_{\alpha/2}\frac{\sigma_x}{\sqrt{n}}$$

A partir de esta expresión, sabiendo que el radio del intervalo es el error muestral máximo que se comete en la estimación, despejando n obtendremos una expresión para el tamaño de muestra necesario en muestreo con reemplazamiento:

$$d_{máx} = Z_{\alpha/2}\frac{\sigma_x}{\sqrt{n}} \Rightarrow$$

$$\Rightarrow n = Z^2_{\alpha/2}\frac{\sigma_x^2}{d^2_{máx}}$$

Esta expresión es la que utilizaremos para calcular el tamaño de muestra, y que **garantizará que, con una probabilidad 1-α, el error muestral máximo que se cometa en la estimación será $d_{máx}$**

En el caso de **muestreo sin reemplazamiento**, el intervalo debe incluir el corrector para poblaciones finitas, y por tanto la expresión de éste varía:

$$\overline{X} \pm Z_{\alpha/2}\frac{\sigma_x}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

Al igual que en el caso anterior, sabiendo que el radio del intervalo expresa el error muestral máximo que se comete en la estimación, operamos para despejar n:

$$d_{máx} = Z_{\alpha/2}\frac{\sigma_x}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} \Rightarrow d_{máx} = Z_{\alpha/2}\sigma_x\sqrt{\frac{N-n}{n(N-1)}} \Rightarrow$$

$$\Rightarrow\sqrt{\frac{N-n}{n(N-1)}} = \frac{d_{máx}}{Z_{\alpha/2}\sigma_x} \Rightarrow \frac{N-n}{n(N-1)} = \frac{d^2_{máx}}{Z^2_{\alpha/2}\sigma_x^2} \Rightarrow$$

$$\Rightarrow N-n = n(N-1)\frac{d^2_{máx}}{Z^2_{\alpha/2}\sigma_x^2} \Rightarrow n\left[1+(N-1)\frac{d^2_{máx}}{Z^2_{\alpha/2}\sigma_x^2}\right] = N \Rightarrow$$

$$\Rightarrow n = \frac{N}{1+(N-1)\dfrac{d^2_{máx}}{Z^2_{\alpha/2}\sigma_x^2}} = \frac{N}{1+N\dfrac{d^2_{máx}}{Z^2_{\alpha/2}\sigma_x^2}-\dfrac{d^2_{máx}}{Z^2_{\alpha/2}\sigma_x^2}} =$$

$$= \frac{Z^2_{\alpha/2}\dfrac{\sigma_x^2}{d^2_{máx}}}{\dfrac{Z^2_{\alpha/2}\dfrac{\sigma_x^2}{d^2_{máx}}+N-1}{N}} = \frac{Z^2_{\alpha/2}\dfrac{\sigma_x^2}{d^2_{máx}}}{1+\dfrac{Z^2_{\alpha/2}\dfrac{\sigma_x^2}{d^2_{máx}}-1}{N}}$$

Es decir, poniendo en relación la fórmula para calcular el tamaño de muestra para un MSR con la fórmula para calcular el tamaño de muestra en MCR llegamos a la siguiente expresión:

$$n_{sr} = \frac{n_{cr}}{1+\dfrac{n_{cr}-1}{N}}$$

Un aspecto esencial es entonces analizar de qué depende el tamaño de la muestra a analizar para construir un intervalo bajo las exigencias planteadas previamente. Podemos decir que dicho tamaño depende:

** Del nivel de confianza $(1-\alpha)$ que deseamos para el intervalo, en una relación directa.

** Del error muestral máximo ($d_{máx}$) que se está dispuesto a cometer en la estimación, en una relación inversa.

** De la dispersión (σ^2) de la variable que se estudia en la población de partida, en una relación directa.

** También (pero muy poco) del tamaño de la población, en una relación directa.

Ejemplo: Calcular el tamaño de muestra necesario para construir un intervalo de confianza para la media de una población, con un nivel de confianza del 90% y un error muestral máximo en la estimación de 2 unidades, sabiendo que la dispersión de la variable de partida (la varianza de la variable de partida) es de 40 unidades^2 y que el muestreo se llevará a cabo con reemplazamiento.

Aplicando la formula deducida:

$$n = Z^2_{\alpha/2} \frac{\sigma_x^2}{d^2_{máx}} =$$

$$= 1,645^2 \frac{40}{2^2} = 27,06 \approx 28$$

Supongamos ahora que deseamos que el nivel de confianza aumente hasta el 99% manteniendo todo lo demás constante. Intuitivamente podemos suponer que para tener una mayor confianza de encerrar la verdadera media poblacional en un intervalo del mismo radio necesitamos un número mayor de observaciones. Veamos lo que sucede aplicando la fórmula:

$$n = 2,575^2 \frac{40}{2^2} = 66,30 \approx 67$$

Vemos por tanto que aumentar el nivel de confianza manteniendo todos los demás factores constantes nos obliga a aumentar el tamaño de la muestra.

Supongamos ahora que manteniendo el nivel de confianza del 90% deseamos que el error muestral máximo disminuya de 2 unidades a 1 unidad. Por pura lógica podemos deducir que para tener la misma probabilidad de encerrar la media poblacional en un intervalo de menor tamaño necesitaremos un mayor número de observaciones. Comprobémoslo:

$$n = 1,645^2 \frac{40}{1^2} = 108,241 \approx 109$$

Podemos afirmar por tanto que disminuir el error muestral máximo que se desea cometer supone aumentar el tamaño de la muestra. De hecho, se puede comprobar fácilmente que reducir a la mitad el error muestral (pasar de d a d/2) supone cuadruplicar el tamaño de la muestra.

Supongamos ahora que la varianza de la variable de partida no es de 40 unidades² sino de 100 unidades². Calculemos el tamaño de muestra necesario para un nivel de confianza del 90% y un error muestral máximo de 2 unidades:

$$n = 1,645^2 \frac{100}{2^2} = 67,65 \approx 68$$

Esto nos lleva a decir que el tamaño de muestra depende de la dispersión de la variable de partida. Las poblaciones muy homogéneas necesitarán tamaños de muestra reducidos mientras que las poblaciones heterogéneas (esto es, con una gran varianza) precisarán de tamaños de muestra mayores para alcanzar los mismos objetivos de seguridad y precisión (nivel de confianza y error muestral máximo).

Por último, planteemos ahora que el muestreo no se lleva a cabo con reemplazamiento sino que por el contrario se muestrea sin reemplazamiento y la población es finita y de tamaño conocido.

Calculemos el tamaño de muestra necesario para que el nivel de confianza sea del 90%, el error muestral máximo sea de 2 unidades y para una población de tamaño N=300:

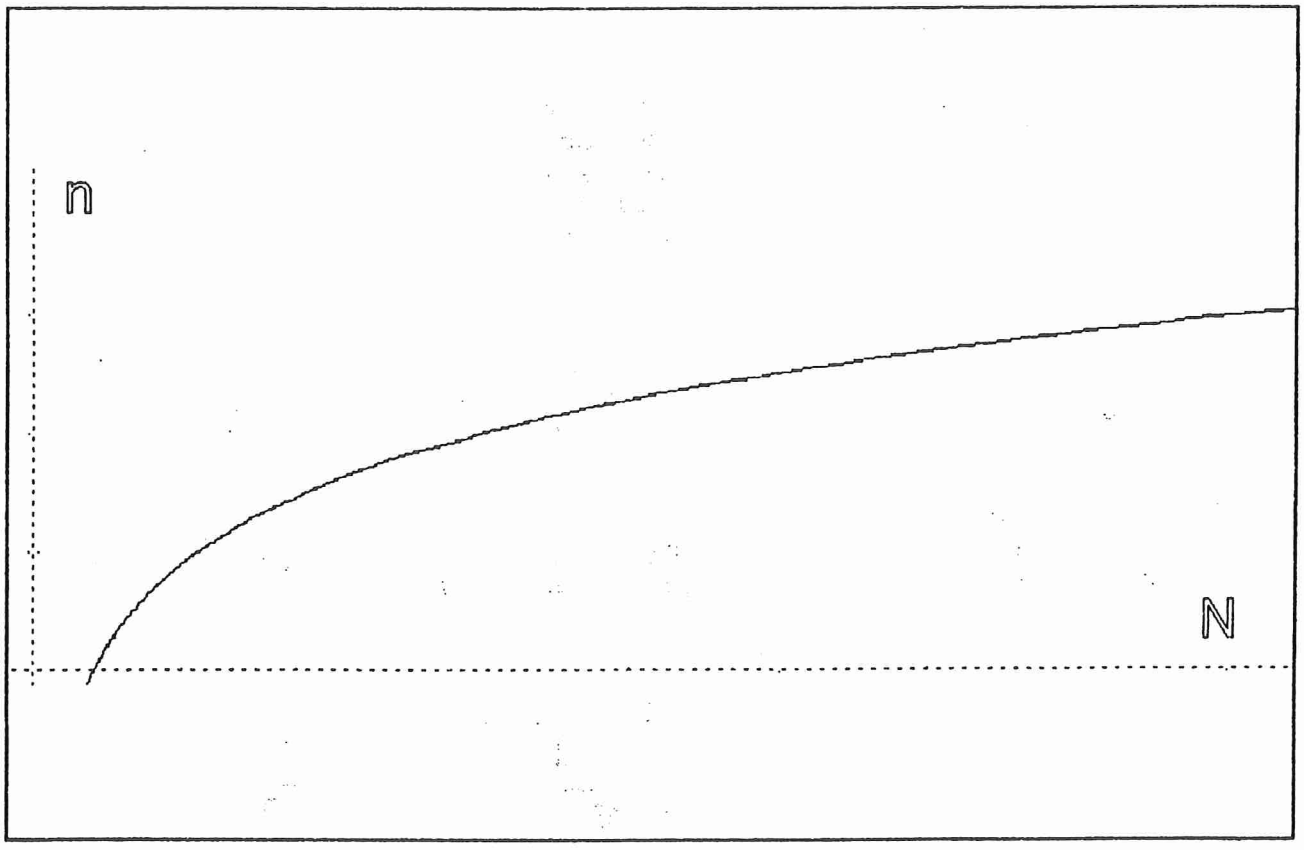
$$n_{sr} = \frac{n_{cr}}{1 + \frac{n_{cr}-1}{N}} = \frac{28}{1 + \frac{27}{300}} \approx 26$$

En el caso de que N sea igual a 1.000.000:

$$n_{sr} = \frac{28}{1+\frac{27}{1000000}} \approx 28$$

Como vemos, el tamaño de la población afecta muy poco al tamaño de muestra necesario para cumplir los requerimientos de nivel de confianza y error muestral máximo dada una cierta varianza de la población de partida. Debemos desterrar la idea de que una población de tamaño doble requiere un tamaño de muestra doble. **La creencia muy extendida de que poblaciones grandes requieren tamaños de muestra grandes no es rigurosamente cierta.** Lo que sí afecta es la dispersión de la población: poblaciones dispersas (muy heterogéneas) sí requieren tamaños de muestra más grandes.

El efecto del tamaño de la población sobre el de la muestra podemos representarlo sobre una gráfica del siguiente tipo:



Ejemplo: Un director de personal desea estimar la calificación media obtenida en un test al que se han presentado un millón de candidatos a puestos de trabajo. Quiere que la estimación tenga un error máximo de dos puntos y desea un nivel de confianza del 95% ¿Qué tamaño de muestra tendrá que analizar?

X: Calificación obtenida en el test por 1 candidato.

Nat X --> ? σ^2= ? μ_x = ? (es lo que deseamos estimar)

$d_{máx} = 2$ $1-\alpha = 0,95$

Aplicando la fórmula deducida:

$$n_{cr} = Z^2_{\alpha/2} \frac{\sigma^2_x}{d^2_{máx}} =$$

$$= 1,96^2 \frac{?}{2^2}$$

Como vemos, no podemos calcular directamente el tamaño de muestra necesario para cumplir los requerimientos **ya que desconocemos la varianza de la población de partida.** Sin embargo, tenemos varias alternativas para aproximarnos al valor de la varianza de la población de partida:

* Basándonos en estudios previos sobre esa población podemos conocer aproximadamente cuál es el valor de la varianza. En el caso de optar por esta alternativa es sumamente importante que el estudio previo se haya realizado sobre esa misma variable y que no exista un gran desfase temporal entre ambos estudios ya que el paso del tiempo puede desvirtuar la validez del valor supuesto para la varianza.

* Realización de una muestra piloto, esto es, llevar a cabo una muestra no aleatoria de pequeño tamaño que sirva como aproximación a la población, de modo que nos sirva para probar o "pilotar" cuestiones como el diseño del cuestionario, el tiempo que nos va a llevar hacer una encuesta, el coste de cada entrevista, la tasa de rechazos en las respuestas y también **para hacer aproximaciones básicas al valor de los parámetros de la población de partida.**

* Nos podemos servir del recorrido de la variable. El recorrido de la variable es la diferencia entre los valores máximo y mínimo que puede tomar esa variable. Si la población de

partida es de naturaleza normal, sabiendo que en una distribución normal entre la media menos tres desviaciones estándar y la media más tres desviaciones estándar se encuentran aproximadamente el 100% de las observaciones, podemos afirmar que el recorrido de la variable es aproximadamente igual a seis desviaciones estándar:

$$P\ (\mu_x - 3\sigma_x \leq X \leq \mu_x + 3\sigma_x) \approx 1 \Rightarrow$$

$$\Rightarrow R \approx 6\sigma_x \Rightarrow \sigma_x \approx \frac{R}{6} \Rightarrow \sigma^2{}_x \approx \left(\frac{R}{6}\right)^2$$

Si la variable no es normal, podemos llevar a cabo una aproximación de la varianza de la población de partida mediante la siguiente ficción:

Supongamos que disponemos de una población con sólo dos observaciones a y b. Evidentemente, ésta es la población más dispersa que podemos imaginar (solamente dos valores extremos y ningún valor central). Calculemos la varianza para esta población:

$$\sigma_x^2 = \frac{(a-\mu_x)^2 + (b-\mu_x)^2}{2} = \frac{\left(\frac{R}{2}\right)^2 + \left(\frac{R}{2}\right)^2}{2} = \left(\frac{R}{2}\right)^2$$

Dado que esa población con sólo dos observaciones es la más dispersa que podemos imaginar, podemos afirmar que ninguna varianza de ninguna variable puede superar el cuadrado de la mitad de su rango. O dicho con otras palabras, la desviación estándar de cualquier variable será como máximo igual a la mitad del rango.

Vemos que en todo caso, en la estimación de la varianza de la variable de partida, estamos tratando de obtener valores por exceso. Es preferible obtener un valor por exceso de la varianza de la población de partida que un valor por defecto.

Si sobrevaloramos la varianza de la población de partida, el tamaño de muestra que calcularemos será superior al que efectivamente necesitaríamos para cumplir los requerimientos. Por lo tanto, cumpliremos lo que se nos pide sobradamente. Por el contrario, si infravaloramos la varianza de la población de partida, el tamaño de muestra que tomaremos será inferior al que

efectivamente necesitaríamos. Esto nos llevará a no cumplir los requerimientos. Por tanto, trataremos de estimar la varianza actuando por exceso, aún sabiendo que por ello incurriremos en unos costes mayores.

Supongamos que ante el desconocimiento de la varianza de la variable de partida, el director de personal decide estimarla basándose en los 20 primeros exámenes. Toma los 20 primeros exámenes como muestra piloto y observa que la varianza en la puntuación es de 20^2 puntos^2. Con este dato, podemos calcular el tamaño de muestra necesario para estimar la media poblacional:

$$n_{cr} = Z^2_{\alpha/2} \frac{\hat{\sigma}_x^2}{d^2_{máx}} = 1,96^2 \frac{20^2}{2^2} \approx 385$$

$$n_{sr} = \frac{n_{cr}}{1+\frac{n_{cr}-1}{N}} = \frac{385}{1+\frac{384}{1000000}} \approx 385$$

Continuando con el ejemplo, supongamos que una vez tomada la muestra de 385 observaciones obtenemos en la muestra que la media muestral vale 75,6 puntos y que la varianza muestral vale $14,7^2$ puntos^2. Construyamos el intervalo de confianza para la estimación de la media poblacional. El intervalo viene dado por la siguiente expresión:

$$\overline{X} \pm Z_{\alpha/2} \frac{S}{\sqrt{n}} \sqrt{1-\frac{n}{N}} =$$

$$= 75,6 \pm 1,96 \frac{14,7}{\sqrt{385}} \sqrt{1-\frac{385}{1000000}} =$$

$$75,6 \pm 1,47$$

$$P\,(74,13 \leq \mu_x \leq 77,07) = 0,95$$

No podemos afirmar con absoluta certeza que la verdadera puntuación media del millón de candidatos se encuentre entre 74,13 y 77,07 (el nivel de confianza no es 1)

Lo que si podemos afirmar con una probabilidad de acierto del 95% es que, analizando muestras de 385 tests, la máxima diferencia entre la media muestral y la verdadera media poblacional es como mucho de 1,47 puntos.

Como podemos ver, hemos alcanzado un error muestral máximo de 1,47 puntos (error máximo a posteriori) cuando el error muestral máximo que estábamos dispuestos a aceptar (error máximo a priori) era de 2 puntos.

Esto se debe a que la varianza muestral de la muestra de tamaño 385 ha sido de $14,7^2$ puntos^2 mientras que la varianza muestral que hemos tomado para calcular el tamaño de muestra necesario había sido de 20^2 puntos^2 (estaba sobrevalorada). Evidentemente es mucho más fiable la estimación basada en una muestra de tamaño 385 que en una muestra de tamaño 20.

En este caso, la varianza muestral ha resultado ser menor de lo que en un principio esperábamos y eso ha redundado en un menor error muestral máximo.

El hecho de que el error máximo a posteriori sea inferior al error máximo a priori puede parecer una situación deseable. Hay que señalar que no se trata de una situación "tan deseable", ya que supone haber realizado un número de observaciones superior al que necesitábamos para el error máximo que estábamos dispuestos a admitir, lo que supone un aumento del coste del estudio.

Si hubiéramos conocido previamente que la varianza no era 20^2 puntos^2 sino $14,7^2$ puntos^2, habríamos necesitado tan solo 208 observaciones frente a las 385 que hemos realizado. Esto evidentemente supone un coste añadido que tenemos que afrontar. El problema es que sólo podemos conocer el valor de la varianza a posteriori, esto es, después de efectuar el muestreo.

Supongamos ahora que una vez realizadas las 385 observaciones llegamos a que el valor de la varianza muestral es de 25^2 puntos^2 en vez de los 20^2 puntos^2 que habíamos estimado con la muestra piloto.

En ese caso el intervalo de confianza para el parámetro media poblacional vendría dado por:

$$75,6 \pm 1,96 \frac{25}{\sqrt{385}} \sqrt{1 - \frac{385}{1000000}} =$$

$$= 75,6 \pm 2,496$$

$$P\,(73,104 \leq \mu_x \leq 78,096) = 0,95$$

Analizando muestras de 385 tests, la máxima diferencia que podemos observar entre la media muestral y la media poblacional (error muestral máximo) es de 2,49 puntos. Esto es cierto con una probabilidad del 95% (nivel de confianza).

Como podemos ver, el error muestral máximo a priori (2, con el que habíamos calculado n) se ha visto superado por el error máximo a posteriori (2,496 el que se ha obtenido en el intervalo). Por tanto, no hemos cumplido los requerimientos de precisión que se nos exigían.

Esto se debe a que la estimación inicial en la muestra piloto de la varianza de la población de partida no ha sido la adecuada, estaba infravalorada. Para conseguir un error muestral máximo de 2 puntos con una varianza poblacional de 25^2 puntos^2 necesitaríamos 601 encuestas.

> **Esto nos lleva a concluir que el cálculo o la estimación del valor de la varianza de la población de partida no es un tema baladí.** Debe dedicarse un cierto esfuerzo para estimar adecuadamente el valor de la varianza de la población de partida ya que ello puede evitar incurrir en otro tipo de costes.

¿Qué es mejor, pasarse o no llegar en la estimación de la varianza? Lo mejor es tratar de determinar correctamente la varianza.

D. Análisis que involucran a dos poblaciones

- MUESTRAS INDEPENDIENTES

Sean dos muestras aleatorias

$$X_1, X_2, \ldots, X_{n_x}$$

$$Y_1, Y_2, \ldots, Y_{n_y}$$

extraídas de dos poblaciones en las que se estudian sendas variables normales e independientes (X e Y), con medias

$$\mu_x \quad y \quad \mu_y$$

y varianzas

$$\sigma_x^2 \quad y \quad \sigma_y^2$$

respectivamente. Las medias poblacionales son desconocidas y queremos estimarlas. Más exactamente queremos llevar a cabo una estimación para la diferencia de las medias poblacionales:

$$\mu_x - \mu_y$$

Pueden plantearse varias situaciones según conozcamos o no las varianzas poblacionales:

(a) Caso de varianzas poblacionales conocidas:

$$X \approx N(\mu_x; \sigma_x^2) \qquad Y \approx N(\mu_y; \sigma_y^2)$$

Sabemos entonces que:

$$\overline{X} \approx N(\mu_x; \frac{\sigma_x^2}{n_x}) \qquad \overline{Y} \approx N(\mu_y; \frac{\sigma_y^2}{n_y})$$

POR TANTO:

$$\overline{X}-\overline{Y} \approx N(\mu_x-\mu_y; \frac{\sigma_x^2}{n_x}+\frac{\sigma_y^2}{n_y})$$

Tipificando la diferencia de las distribuciones muestrales de medias tendremos:

$$\frac{\overline{X}-\overline{Y} - (\mu_x-\mu_y)}{\sqrt{\frac{\sigma_x^2}{n_x}+\frac{\sigma_y^2}{n_y}}} = Z$$

De esta forma, dado un nivel de confianza $(1-\alpha)$, podremos encontrar en las tablas de Z el valor

$$Z_{\alpha/2}$$

que cumpla que

$$P\ (-Z_{\alpha/2} \leq Z \leq Z_{\alpha/2}) = 1-\alpha$$

En la afirmación anterior, mediante la sustitución de la expresión Z por una equivalente (la tipificación de la diferencia de distribuciones muestrales de medias) tendremos que:

$$P\ (-Z_{\alpha/2} \leq \frac{\overline{X}-\overline{Y} - (\mu_x-\mu_y)}{\sqrt{\frac{\sigma_x^2}{n_x}+\frac{\sigma_y^2}{n_y}}} \leq Z_{\alpha/2}) = 1-\alpha \Rightarrow$$

$$P\ (\overline{X}-\overline{Y} - Z_{\alpha/2}\sqrt{\frac{\sigma_x^2}{n_x}+\frac{\sigma_y^2}{n_y}} \leq \mu_x-\mu_y \leq \overline{X}-\overline{Y} + Z_{\alpha/2}\sqrt{\frac{\sigma_x^2}{n_x}+\frac{\sigma_y^2}{n_y}}) = 1-\alpha$$

Por tanto, concluímos afirmando que en el caso de que las varianzas poblacionales sean conocidas, el intervalo de confianza para el parámetro diferencia de medias poblacionales viene dado por la siguiente expresión:

$$CENTRO \pm RADIO$$

$$(\overline{X}-\overline{Y}) \pm Z_{\alpha/2}\sqrt{\frac{\sigma_x^2}{n_x}+\frac{\sigma_y^2}{n_y}}$$

b) Caso de varianzas poblacionales desconocidas, pero donde la prueba de homocedasticidad indica que son iguales:

Recordemos que las varianzas poblacionales se estiman conjuntamente, usando la información de ambas muestras. Tendremos entonces que:

$$\overline{X}-\overline{Y} \approx N(\mu_x-\mu_y; \frac{S_p^2}{n_x}+\frac{S_p^2}{n_y})$$

$$Donde\ S^2_p = \frac{(n_x-1)S^2_x + (n_y-1)S^2_y}{n_x+n_y-2}$$

Tipificando la diferencia de las distribuciones muestrales de medias tendremos:

$$\frac{\overline{X}-\overline{Y} - (\mu_x-\mu_y)}{\sqrt{\frac{S_p^2}{n_x}+\frac{S_p^2}{n_y}}} = t_{n_x+n_y-2}$$

De esta forma, dado un nivel de confianza (1-α), podremos encontrar en las tablas de la t Student el valor

$$t_{n_x+n_y-2,\alpha/2}$$

que cumpla que

$$P\ (-t_{n_x+n_y-2,\alpha/2} \le t_{n_x+n_y-2} \le t_{n_x+n_y-2,\alpha/2}) = 1-\alpha$$

En la afirmación anterior, mediante la sustitución de la expresión central t por una equivalente (la tipificación de la diferencia de distribuciones muestrales de medias) tendremos que:

$$P(-t_{n_x+n_y-2,\alpha/2} \le \frac{\overline{X}-\overline{Y} - (\mu_x-\mu_y)}{\sqrt{\frac{S^2_p}{n_x}+\frac{S^2_p}{n_y}}} \le t_{n_x+n_y-2,\alpha/2}) = 1-\alpha \rightarrow$$

$$P(\overline{X}-\overline{Y} - t_{n_x+n_y-2,\alpha/2}\sqrt{\frac{S^2_P}{n_x}+\frac{S^2_P}{n_y}} \leq \mu_x-\mu_y \leq \overline{X}-\overline{Y} + t_{n_x+n_y-2,\alpha/2}\sqrt{\frac{S^2_P}{n_x}+\frac{S^2_P}{n_y}}) = 1-\alpha$$

Por tanto, concluímos afirmando que en el caso de que las varianzas poblacionales sean desconocidas pero iguales, el intervalo de confianza para el parámetro diferencia de medias poblacionales viene dado por la siguiente expresión:

$$(\overline{X}-\overline{Y}) \pm t_{n_x+n_y-2,\alpha/2}\sqrt{\frac{S^2_P}{n_x}+\frac{S^2_P}{n_y}}$$

c) Caso de varianzas poblacionales desconocidas y donde la prueba de homocedasticidad indica además que son distintas.

Recordemos que las varianzas poblacionales se estiman por separado, usando la información de cada muestra. Tendremos entonces que:

$$\overline{X}-\overline{Y} \approx N(\mu_x-\mu_y; \frac{S_x^2}{n_x}+\frac{S_y^2}{n_y})$$

Tipificando la diferencia de las distribuciones muestrales de medias tendremos:

$$\frac{(\overline{X}-\overline{Y})-(\mu_x-\mu_y)}{\sqrt{\frac{S_x^2}{n_x}+\frac{S_y^2}{n_y}}} = t_k$$

con grados de libertad:

$$k = \frac{(n_x-1)(n_y-1)}{(n_y-1)c^2+(1-c)^2(n_x-1)}$$

$$donde\ c = \frac{\frac{S^2_x}{n_x}}{\frac{S^2_x}{n_x} + \frac{S^2_y}{n_y}}$$

En la expresión anterior, si los grados de libertad k resultan no ser un número entero, se suelen redondear por defecto al entero más próximo antes de consultar las tablas de la t Student.

De esta forma, dado un nivel de confianza $(1-\alpha)$, podremos encontrar en las tablas de la t Student el valor

$$t_{k,\alpha/2}$$

que cumpla que

$$P\ (-t_{k,\alpha/2} \leq t_k \leq t_{k,\alpha/2}) = 1-\alpha$$

En la afirmación anterior, mediante la sustitución de la expresión central t por una equivalente (la tipificación de la diferencia de distribuciones muestrales de medias) tendremos que:

$$P(-t_{k,\alpha/2} \leq \frac{\overline{X}-\overline{Y} - (\mu_x-\mu_y)}{\sqrt{\frac{S^2_x}{n_x} + \frac{S^2_y}{n_y}}} \leq t_{k,\alpha/2}) = 1-\alpha \rightarrow$$

$$P(\overline{X}-\overline{Y} - t_{k,\alpha/2}\sqrt{\frac{S^2_x}{n_x}+\frac{S^2_y}{n_y}} \leq \mu_x-\mu_y \leq \overline{X}-\overline{Y} + t_{k,\alpha/2}\sqrt{\frac{S^2_x}{n_x}+\frac{S^2_y}{n_y}}) = 1-\alpha$$

Por tanto, concluímos afirmando que en el caso de que las varianzas poblacionales sean desconocidas y además diferentes, el intervalo de confianza para el parámetro diferencia de medias poblacionales viene dado por la siguiente expresión:

$$(\overline{X}-\overline{Y}) \pm t_{k,\alpha/2}\sqrt{\frac{S^2_x}{n_x}+\frac{S^2_y}{n_y}}$$

MUESTRAS DEPENDIENTES O RELACIONADAS

Partimos de dos poblaciones en las que se estudian dos variables aleatorias normales, respectivamente X e Y.

y si no lo suponemos.

De una y sólo de una de las dos poblaciones extraemos una muestra al azar:

$$X_1, X_2, \ldots, X_n$$

Esta muestra extraída al azar de la primera población determina inmediatamente las observaciones a extraer de la segunda población

$$X_1 \rightarrow Y_1$$
$$X_2 \rightarrow Y_2$$
$$\ldots\ldots$$
$$X_n \rightarrow Y_n$$

La muestra de la segunda población no se extrae al azar, sino que depende totalmente de la muestra extraída de la primera población. Es por ello que hablamos de muestras dependientes o relacionadas.

La forma de resolver este tipo de problemas consiste en reducir el número de variables con las que trabajamos. Para ello, definiremos una nueva variable D que será la diferencia de las variables de origen

$$D = X - Y$$

De este modo, en las muestras pasaremos de tener 2n observaciones a tener simplemente n diferencias o pares, a saber:

$$\begin{aligned} d_1 &= X_1 - Y_1 \\ d_2 &= X_2 - Y_2 \\ &\cdots\cdots\cdots \\ d_n &= X_n - Y_n \end{aligned}$$

Dado que las dos variables de origen X e Y tenían naturaleza normal, la nueva variable D que hemos definido tendrá también naturaleza normal, siendo su valor esperado:

$$E(D) = E(X-Y) = E(X) - E(Y) = \mu_x - \mu_y$$

Su varianza sin embargo será siempre desconocida, al serlo también las varianzas de las variables X e Y de partida

$$\sigma_d^2 = ?$$

A partir de la muestra de n diferencias o pares que hemos obtenido de las poblaciones de partida, definimos los siguientes estadísticos muestrales:

$$\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n}$$

$$S^2_d = \frac{\sum_{i=1}^{n} (d_i - \bar{d})^2}{n-1}$$

Esta variable $\bar{d}$ que hemos definido será la media muestral de las diferencias, y de ella podemos decir que:

$$Nat(\overline{d}) \rightarrow NORMAL$$

$$E(\overline{d}) = \mu_d = \mu_x - \mu_y$$

$$Var(\overline{d}) = \frac{\sigma^2_d}{n} = ? : \frac{\sigma_x^2 + \sigma_y^2}{n \cdot n}$$

Si las variables son dependientes, esto no se puede calcular, así que olvidarnos de hacerlo así.

$$\hat{Var}(\overline{d}) = \frac{S^2_d}{n}$$

Basándonos en esto, podemos afirmar que:

$$\frac{\overline{d} - \mu_d}{\frac{S_d}{\sqrt{n}}} = t_{n-1} \Rightarrow \frac{\overline{d} - (\mu_x - \mu_y)}{\frac{S_d}{\sqrt{n}}} = t_{n-1}$$

De esta forma, dado un nivel de confianza $(1-\alpha)$, podremos encontrar en las tablas de la t Student el valor

$$t_{n-1;\alpha/2}$$

que cumpla que

$$P\left(-t_{n-1,\alpha/2} \leq t_{n-1} \leq t_{n-1,\alpha/2}\right) = 1-\alpha$$

En la afirmación anterior, mediante la sustitución de la expresión t por una equivalente (la tipificación de la media muestral de pares) tendremos que:

$$P\left(-t_{n-1;\alpha/2} \leq \frac{\overline{d} - (\mu_x - \mu_y)}{S_d/\sqrt{n}} \leq t_{n-1;\alpha/2}\right) = 1-\alpha \Rightarrow$$

$$P\left(\overline{d} - t_{n-1;\alpha/2}\frac{S_d}{\sqrt{n}} \leq \mu_x - \mu_y \leq \overline{d} + t_{n-1;\alpha/2}\frac{S_d}{\sqrt{n}}\right) = 1-\alpha$$

Por tanto, concluímos afirmando que en el caso de que las muestras se hayan obtenido de forma dependiente o relacionada, el intervalo de confianza para el parámetro diferencia de medias poblacionales viene dado por la siguiente expresión

CENTRO ± RADIO
$\bar{d} \pm t_{n-1,\alpha/2} \frac{S_d}{\sqrt{n}}$

En el caso de que el muestreo se lleve a cabo sin reemplazamiento debe incluirse en el intervalo el corrector para poblaciones finitas:

$$\sqrt{1-\frac{n}{N}}$$

Ejemplo Supongamos que queremos saber si, por término medio, en las parejas en que ambos cónyuges trabajan, los hombres ganan más que las mujeres. Para ello seleccionamos al azar a 15 parejas (hombre-mujer) en las que ambos trabajan y les preguntamos cuál es su sueldo.

Los datos obtenidos son los siguientes: (100,90) (104,102) (108,110) (92,85) (80,85) (76,70) (125,130) (130,110) (110,106) (100,94) (84,86) (96,90) (108,108) (78,75) (96,92). Al nivel de confianza del 95% construir un intervalo para la diferencia media del sueldo de los hombres respecto a las mujeres.

X: Salario percibido por 1 hombre que trabaja.
Y: Salario percibido por 1 mujer que trabaja.

Desconocemos la naturaleza de ambas variables pero debemos suponerlas normales. También desconocemos el valor de los parámetros varianzas y medias poblacionales.

Es evidente que se trata de una prueba de muestras dependientes ya que una vez elegido un miembro de la pareja el otro está inequívocamente determinado. Si elegimos un hombre, el miembro de la otra población que pasará a formar parte de la otra muestra será su esposa. Si elegimos una mujer, la observación en la otra población corresponderá al marido.

Definamos la nueva variable aleatoria d: "Diferencia entre el sueldo de un hombre y una mujer (su pareja) que trabajan".

$$d = X - Y$$

$$d \rightarrow NORMAL\ (\mu_d\ ;\ \sigma^2{}_d)$$

$$\mu_d = \mu_x - \mu_y\ (lo\ que\ queremos\ estimar)$$

$$\sigma^2{}_d = ?$$

Nos encontramos por tanto ante el caso de una única variable (d), con varianza poblacional desconocida y en la que deseamos construir un intervalo de confianza para su parámetro media poblacional.

Para ello contamos con una muestra de tamaño 15. Los valores de la variable d que se obtienen en la muestra son:

100- 90 = 10 ; 104-102 = 2 ; 108-110 = -2 ; 92-85 = 7

80- 85 = -5 ; 76- 70 = 6 ; 125-130 = -5 ; 130-110= 20

110-106 = 4 ; 100- 94 = 6 ; 84- 86 = -2 ; 96- 90 = 6

108-108 = 0 ; 78- 75 = 3 ; 96- 92 = 4

$$\overline{d} = 3,6$$

$$\hat{\sigma}^2{}_d = s^2{}_d = 6,356^2$$

El intervalo para el parámetro buscado vendrá dado por la siguiente expresión:

$$\overline{d} \pm t_{n-1;\alpha/2}\frac{S_d}{\sqrt{n}} = 3,6 \pm 2,1448\frac{6,356}{\sqrt{15}} =$$

$$= 3,6 \pm 3,5 \Rightarrow$$

$$\Rightarrow P\ (0,1 \leq \mu_d \leq 7,1) = 0,95$$

$$\Rightarrow P\ (0,1 \leq \mu_x - \mu_y \leq 7,1) = 0,95$$

Esto es, podemos afirmar con una probabilidad del 95% que la verdadera diferencia media entre los sueldos de los hombres y las mujeres que trabajan (parámetro) no se separa en más de 3,5 unidades monetarias (error muestral máximo) de la diferencia media obtenida en la muestra (estimador).

Aunque no sabemos en qué punto del intervalo pueda estar encerrado el parámetro buscado, dado que ambos extremos del intervalo son positivos, podemos afirmar con probabilidad del 95% que los hombres que trabajan ganan, por término medio, más que las mujeres que trabajan.

VII. ESTIMACION POR INTERVALOS DE LA PROPORCION POBLACIONAL DE EXITOS

A. Condiciones de aproximación a la distribución normal

Recordemos que en esta ocasión el punto de partida es una población en la que se estudia una variable cualitativa dicotómica o variable de Boole. Esta variable tan sólo puede tomar dos modalidades de aparición: éxito o fracaso.

De esa población extraemos mediante muestreo aleatorio simple muestras de n observaciones y definimos en las mismas una nueva variable $\bar{p}$, que expresará **la proporción de éxitos en la muestra de tamaño n.** Como sabemos ya,

$$E(\bar{p}) = P$$

$$Var(\bar{p}) = \frac{PQ}{n} \quad (MCR) \qquad 1$$

$$= \frac{PQ}{n} \frac{N-n}{N-1} \quad (MSR) \qquad 2$$

Sabemos también que para valores de n suficientemente grandes la distribución muestral de proporciones se comporta según un modelo normal, y que su tipificación dará una variable Z.

$$\frac{\bar{p} - P}{\sqrt{\frac{PQ}{n}}} = Z$$

De esta forma, dado un nivel de confianza $(1-\alpha)$, podremos encontrar en las tablas de Z el valor

$$Z_{\alpha/2}$$

que cumpla que

$$P\ (-Z_{\alpha/2} \leq Z \leq Z_{\alpha/2}) = 1-\alpha$$

En la afirmación anterior, mediante la sustitución de la expresión Z por una equivalente (la tipificación de la distribución muestral de la proporción) tendremos que:

$$P\ (-Z_{\alpha/2} \leq \frac{\bar{p}-P}{\sqrt{\frac{PQ}{n}}} \leq Z_{\alpha/2}) = 1-\alpha$$

$$P\ (\bar{p} - Z_{\alpha/2}\sqrt{\frac{PQ}{n}} \leq P \leq \bar{p} + Z_{\alpha/2}\sqrt{\frac{PQ}{n}}) = 1-\alpha$$

Sin embargo, este razonamiento nos lleva a un absurdo, **ya que el intervalo incluye parámetros desconocidos:** para estimar P necesitamos conocer PQ.

Para resolver el problema con que nos encontramos creamos la siguiente ficción: en la variable de Boole de partida asignamos al éxito el valor 1 y al fracaso el valor 0, y calculamos la media y la varianza de la variable así definida. Observamos que **la media de esa nueva variable definida coincide con la proporción poblacional de éxitos P y que la varianza de la nueva variable coincide con el producto PQ.**

variable dicotómica, sólo 2 posibles resultados.

si no se conoce la varianza poblacional, sustituimos ésta por la varianza muestral: $\hat{S}^2 = S^2 \frac{n}{n-1} = pq\frac{n}{n-1} =$

Comprobémoslo con un ejemplo: Sea una población de tamaño 10 en la que tenemos 3 éxitos y 7 fracasos. Asignemos al éxito el valor 1 y al fracaso el valor 0. Nuestra población estará ahora formada por los siguientes valores: 1,1,1,0,0,0,0,0,0,0.

Si calculamos la media de esta variable observaremos que es 0,3 y que coincide con la proporción poblacional de éxitos (3/10).

La varianza poblacional de la nueva variable es 0,21 que coincide con el producto PQ (3/10 x 7/10 = 0,21). Esto nos lleva a una clara conclusión: **la proporción de éxitos es una media** de una variable que sólamente toma los valores 0 y 1.

Dado que en el intervalo construído precisamos conocer PQ (la varianza poblacional) para estimar P (la media poblacional), lo que podemos hacer es estimar PQ en base a los datos de la muestra. Como resulta lógico, lo haremos en base a la varianza de la muestra:

si no se conoce la varianza poblacional $\sigma^2 = PQ$

$$\hat{\sigma}^2 = S^2 = \frac{n}{n-1}\,\overline{pq} = \hat{PQ}$$

Por tanto, el intervalo quedará definido del siguiente modo:

$$P\left(\overline{p} - Z_{\alpha/2}\sqrt{\frac{\overline{pq}}{n-1}} \le P \le \overline{p} + Z_{\alpha/2}\sqrt{\frac{\overline{pq}}{n-1}}\right) = 1-\alpha$$

Para MSR incluir en el radio $\sqrt{1-\frac{n}{N}}$

$$\overline{p} \pm Z_{\alpha/2}\sqrt{\frac{\overline{pq}}{n-1}}\sqrt{1-\frac{n}{N}}$$

Ejemplo: El Ayuntamiento de Donostia-San Sebastián desea conocer el porcentaje de ciudadanos que se muestran favorables a la construcción de una nueva plaza de toros. Para ello selecciona al azar y encuesta a 1600 donostiarras, obteniendo los siguientes

resultados: a favor 640, en contra 960. Sabiendo que el censo de donostiarras es aproximadamente de 200.000 nos piden construir un intervalo de confianza al 95% para el porcentaje de donostiarras favorables a la construcción de una nueva plaza de toros. Construir también un intervalo de confianza al 95% para el número total de donostiarras favorables a la construcción de la nueva plaza de toros.

Definimos una variable B del siguiente modo: Una persona entrevistada está de acuerdo (éxito=1) o está en desacuerdo (fracaso=0) con la construcción de una nueva plaza de toros.

Tomamos una muestra de 1600 donostiarras en un muestreo sin reemplazamiento de una población de 200.000 donostiarras y definimos la variable $\overline{p}$ como la **proporción de personas favorables a la construcción de una nueva plaza de toros en 1 muestra de 1600 encuestados.** En concreto, en la muestra que hemos obtenido, el valor de la variable $\overline{p}$ es 640/1600 = 0,4.

El intervalo para la proporción poblacional de éxitos vendrá dado por la siguiente expresión:

$$\overline{p} \pm Z_{\alpha/2}\sqrt{\frac{\overline{pq}}{n-1}}\sqrt{1-\frac{n}{N}} =$$

$$0,4 \pm 1,96\sqrt{\frac{0,4 \times 0,6}{1599}}\sqrt{1-\frac{1600}{200000}} =$$

$$0,4 \pm 0,024$$

$$P\,(0,376 \leq P \leq 0,424) = 0,95$$

Podemos afirmar con probabilidad del 95% que la proporción de ciudadanos donostiarras que son favorables a la construcción de la nueva plaza de toros se encuentra entre el 37,6% y el 42,4% En todo caso, y con probabilidad del 95%,la diferencia máxima entre la auténtica proporción de donostiarras que son favorables a la construcción de la nueva plaza y la proporción obtenida en la muestra será del 2,4%

En lo que se refiere al cálculo del intervalo para el número total de ciudadanos favorables a la construcción de la nueva plaza de toros, no tenemos más que multiplicar ambos extremos del intervalo por el tamaño de la población:

$$P(75200 \leq \begin{matrix} TOTAL \\ DE \\ EXITOS \end{matrix} \leq 84800) = 0,95$$

Todo el desarrollo que hemos hecho se construye sobre la base de que podemos llevar a cabo una aproximación a la normal de la naturaleza hipergeométrica o binomial de la distribución muestral de proporciones. Esto supone que el tamaño de muestra analizado es lo suficientemente grande.

En el caso de que esto no se cumpla deberemos respetar la naturaleza hipergeométrica o binomial de la distribución muestral de proporciones, y recurrir a la figura conocida como ábaco, que nos muestra fácilmente los límites del intervalo de confianza para la proporción poblacional de éxitos dados un cierto nivel de confianza, un cierto número de observaciones y una cierta proporción muestral de éxitos.

B. Cálculo del tamaño de muestra necesario para estimar la proporción poblacional de éxitos

Nos encontramos en la misma situación que cuando tratábamos de calcular el tamaño de muestra necesario para estimar la media poblacional. Una vez que hemos fijado el nivel de confianza que deseamos así como el error muestral máximo que estamos dispuestos a cometer, nos interesa conocer qué tamaño de muestra nos permite satisfacer estos objetivos sabiendo que ambos son contradictorios y sólo se pueden cumplir simultáneamente ampliando el tamaño de la muestra a analizar.

En el caso de muestreo con reemplazamiento tendremos:

$$d_{máx} = Z_{\alpha/2} \sqrt{\frac{PQ}{n}} \Rightarrow$$

$$\Rightarrow n = Z^2_{\alpha/2} \frac{PQ}{d^2_{máx}}$$

Para el caso de muestreo sin reemplazamiento tendremos:

$$d_{máx} = Z_{\alpha/2}\sqrt{\frac{PQ}{n}}\sqrt{\frac{N-n}{N-1}} \Rightarrow d_{máx} = Z_{\alpha/2}\sqrt{PQ}\sqrt{\frac{N-n}{n(N-1)}} \Rightarrow$$

$$\Rightarrow \sqrt{\frac{N-n}{n(N-1)}} = \frac{d_{máx}}{Z_{\alpha/2}\sqrt{PQ}} \Rightarrow \frac{N-n}{n(N-1)} = \frac{d^2_{máx}}{Z^2_{\alpha/2}PQ} \Rightarrow$$

$$\Rightarrow N-n = n(N-1)\frac{d^2_{máx}}{Z^2_{\alpha/2}PQ} \Rightarrow n\left[1+(N-1)\frac{d^2_{máx}}{Z^2_{\alpha/2}PQ}\right] = N \Rightarrow$$

$$\Rightarrow n = \frac{N}{1+(N-1)\frac{d^2_{máx}}{Z^2_{\alpha/2}PQ}} = \frac{N}{1+N\frac{d^2_{máx}}{Z^2_{\alpha/2}PQ}-\frac{d^2_{máx}}{Z^2_{\alpha/2}PQ}} =$$

$$= \frac{Z^2_{\alpha/2}\frac{PQ}{d^2_{máx}}}{\frac{Z^2_{\alpha/2}\frac{PQ}{d^2_{máx}}+N-1}{N}} = \frac{Z^2_{\alpha/2}\frac{PQ}{d^2_{máx}}}{1+\frac{Z^2_{\alpha/2}\frac{PQ}{d^2_{máx}}-1}{N}}$$

Esto es, poniendo en relación los tamaños de muestra para MCR y MSR, llegamos a la siguiente expresión:

$$n_{sr} = \frac{n_{cr}}{1+\frac{n_{cr}-1}{N}}$$

Al igual que sucedía en el cálculo del tamaño de muestra para estimar la media poblacional, la varianza poblacional (en este caso PQ) suele ser desconocida. Tenemos tres alternativas para estimar el valor de PQ:

- Estudios previos.
- Muestra piloto.
- Condiciones desfavorables de muestreo.

Esta última opción (la más frecuentemente utilizada) consiste en encontrar un límite máximo para el valor de PQ. Podríamos decir que se trata de la solución más conservadora, ya que es ofrecer como estimación de la varianza poblacional desconocida su máximo valor. Si seguimos esta opción, nos aseguramos que el intervalo final nunca tendrá un radio (error a posteriori) superior al error máximo admitido a priori al calcular n.

Es evidente que el producto de dos numeros, P y Q, cuya suma debe ser 1 alcanza un valor máximo. Veámoslo:

$$y = PQ = P(1-P)$$

Derivada primera:

$$y' = (1-P) - P = 1-2P$$

Igualando y' *a* $0 \Rightarrow 1-2P = 0 \Rightarrow P = 1/2$

Derivada segunda:

$$y'' = -2 < 0 \Rightarrow P = 1/2 \text{ es máximo}$$

Es decir, cuando P=Q=0,5, el producto PQ alcanza su valor máximo. Dicho con otras palabras, no existe ninguna combinación de dos números menores que 1 cuya suma sea 1 y cuyo producto sea mayor que 0,25. Por tanto, trabajando en condiciones desfavorables de muestreo, la estimación de la varianza poblacional desconocida PQ consistirá en sustituir su valor por 0,25.

Ejemplo: Una compañía de seguros que tiene 9000 asegurados desea calcular qué porcentaje de los mismos va a cambiar de coche en los próximos dos años. Para ello requiere un error máximo del 3% al nivel de confianza del 95% ¿Cuántos asegurados habrá de entrevistar para cumplir estos requerimientos?.

Definamos en primer lugar la variable de Boole de partida:

B: Un asegurado de la compañía va a cambiar (éxito=1) o no va a cambiar (fracaso=0) de coche en los próximos dos años.

Suponemos que se trata de MSR, ya que no estaremos interesados en preguntar dos veces al mismo asociado.

En el 95% de los [illegible]los que construyamos tomando muestras de 955 asociados, el m[illegible]error que podemos cometer al estimar P en base a $\bar{p}$ es del 2,9%

Como vemos, el error muestral máximo a posteriori es menor que el error muestral máximo a priori. Ello se debe a que con las condiciones desfavorables de muestreo nos habíamos "cubierto las espaldas". Posteriormente se ha visto que el producto PQ no era tan grande como 0,25, **aunque no lo sabíamos al calcular el tamaño de la muestra.**

C. Estimaciones que involucran a dos poblaciones con muestras independientes

El punto de partida son dos poblaciones en las que se analizan sendas variables de Boole

B_x *en la primera población*

B_y *en la segunda población*

De cada una de las poblaciones tomamos sendas muestras de gran tamaño

n_x *de la primera población*

n_y *de la segunda población*

Definimos dos nuevas variables aleatorias

$$\bar{p}_x \quad y \quad \bar{p}_y$$

que representarán la proporción muestral de éxitos en la primera y en la segunda muestra.

Sabemos que siempre que los tamaños de muestra sean lo suficientemente grandes se cumplirá que:

Primeramente calcularemos el tamaño de muestra necesario en el caso de muestreo con reemplazamiento:

$$n_{cr} = Z_{\alpha/2}\frac{PQ}{d^2_{máx}} = 1,96^2\frac{PQ}{0,03^2}$$

Dado que carecemos de estudios previos y que tampoco hemos realizado una muestra piloto optaremos por estimar el producto PQ bajo condiciones desfavorables de muestreo:

$$n_{cr} = 1,96^2\frac{0,25}{0,03^2} \approx 1068$$

Para el caso de muestreo sin reemplazamiento tendremos:

$$n_{sr} = \frac{1068}{1+\frac{1067}{9000}} \approx 955$$

Bajo condiciones desfavorables de muestreo, la compañía deberá entrevistar a 955 asociados y tendrá **la absoluta certeza** de que el intervalo que construya no va a tener un error muestral o radio superior al 3% (para un nivel de confianza del 95%)

Supongamos que analizada la muestra, el 60% de los asociados manifiestan estar dispuestos a cambiar de coche en los próximos dos años. El intervalo de confianza vendrá dado por:

$$\overline{p} \pm Z_{\alpha/2}\sqrt{\frac{\overline{pq}}{n-1}}\sqrt{1-\frac{n}{N}} =$$

$$0,6 \pm 1,96 \; x \; 0,0158 \; x \; 0,945 =$$

$$0,6 \pm 0,029$$

$$P\,(0,571 \leq P \leq 0,629) = 0,95$$

$$\overline{p}_x \Rightarrow N\ (P_x\ ;\ \frac{P_xQ_x}{n_x})$$

$$\overline{p}_y \Rightarrow N\ (P_y\ ;\ \frac{P_yQ_y}{n_y})$$

$$\overline{p}_x-\overline{p}_y \Rightarrow N\ (P_x-P_y\ ;\ \frac{P_xQ_x}{n_x}+\frac{P_yQ_y}{n_y})$$

$$\frac{(\overline{p}_x-\overline{p}_y)\ -\ (P_x-P_y)}{\sqrt{\frac{P_xQ_x}{n_x}+\frac{P_yQ_y}{n_y}}} = Z$$

De esta forma, dado un nivel de confianza $(1-\alpha)$, podremos encontrar en las tablas de Z el valor

$$Z_{\alpha/2}$$

que cumpla que

$$P\ (-Z_{\alpha/2} \leq Z \leq Z_{\alpha/2}) = 1-\alpha$$

En la afirmación anterior, mediante la sustitución de la expresión Z por una equivalente (la tipificación de la distribución muestral de la diferencia de proporciones) tendremos que:

$$P\ (-Z_{\alpha/2} \leq \frac{(\overline{p}_x-\overline{p}_y)\ -\ (P_x-P_y)}{\sqrt{\frac{P_xQ_x}{n_x}+\frac{P_yQ_y}{n_y}}} \leq Z_{\alpha/2}) = 1-\alpha \Rightarrow$$

$$P\ [(\overline{p}_x-\overline{p}_y)\ -\ Z_{\alpha/2}\sqrt{\frac{P_xQ_x}{n_x}+\frac{P_yQ_y}{n_y}} \leq P_x-P_y \leq (\overline{p}_x-\overline{p}_y)\ +\ Z_{\alpha/2}\sqrt{\frac{P_xQ_x}{n_x}+\frac{P_yQ_y}{n_y}}) = 1-\alpha$$

Nuevamente nos encontramos con un absurdo, ya que para estimar el parámetro diferencia de proporciones poblacionales necesitamos conocer los correspondientes productos PQ (varianzas poblacionales) en ambas poblaciones. La solución viene por el mismo camino que vimos en el caso de una sola población: estimarlas con las correspondientes varianzas muestrales.

$$\hat{P}_x\hat{Q}_x = \frac{n_x}{n_x-1}\,\overline{p}_x\overline{q}_x$$

$$\hat{P}_y\hat{Q}_y = \frac{n_y}{n_y-1}\,\overline{p}_y\overline{q}_y$$

De este modo, el intervalo para la diferencia de proporciones poblacionales de éxitos quedará expresado del siguiente modo, sin contener ahora información desconocida:

$$P[(\overline{p}_x-\overline{p}_y) \pm Z_{\alpha/2}\sqrt{\frac{p_xq_x}{n_x-1}+\frac{p_yq_y}{n_y-1}}] = 1-\alpha$$

Ejemplo: Una revista quiere saber si porcentualmente es más leída por hombres o por mujeres. Para ello desean construir un intervalo de confianza al 95% para dicha diferencia. Encuestan a 400 hombres al azar y a 600 mujeres también al azar. 100 hombres y 300 mujeres manifestaron ser lectores de la revista. Se pide construir e interpretar el correspondiente intervalo.

Definimos las dos variables de Boole de partida:

B_x : Una mujer encuestada lee (éxito=1) o no lee (fracaso=0) la revista.

B_y: Un hombre encuestado lee (éxito=1) o no lee (fracaso=0) la revista.

n_x= 600 en MSR ---> $\overline{p}_x$= 0,5.

n_y= 400 en MSR ---> $\overline{p}_y$= 0,25.

El intervalo para la diferencia en la proporción de lectores según se trate de hombres o mujeres viene dado por la siguiente expresión:

$$(\overline{p}_x - \overline{p}_y) \pm Z_{\alpha/2}\sqrt{\frac{p_x q_x}{n_x - 1} + \frac{p_y q_y}{n_y - 1}} =$$

$$0,25 \pm 1,96\sqrt{\frac{0,5 \times 0,5}{599} + \frac{0,25 \times 0,75}{399}} =$$

$$0,25 \pm 0,06$$

$$P\ (0,19 \leq P_x - P_y \leq 0,31) = 0,95$$

Esto es, con probabilidad del 95% podemos afirmar que el máximo error que podemos cometer al estimar la auténtica diferencia entre las proporciones de hombres y mujeres lectores de la revista en base a la diferencia observada en dos muestras de tamaño 400 y 600, respectivamente, es de 0,06.

No sabemos en qué punto del intervalo se puede encontrar el parámetro diferencia de proporciones poblacionales de éxitos. Sin embargo, como los dos límites del intervalo tienen signos positivos podremos concluir, con un nivel de confianza del 95% que:

$$P_x > P_y$$

Esto es, de entre las mujeres, el porcentaje de las que lee la revista es mayor que el porcentaje de los que la leen entre los hombres.

VIII. ESTIMACION POR INTERVALOS DE LA VARIANZA POBLACIONAL

Se trata de un caso particular que no sigue la forma general de los intervalos: centro ± radio. Por otra parte, el centro del intervalo no va a ser tampoco el estimador puntual del parámetro.

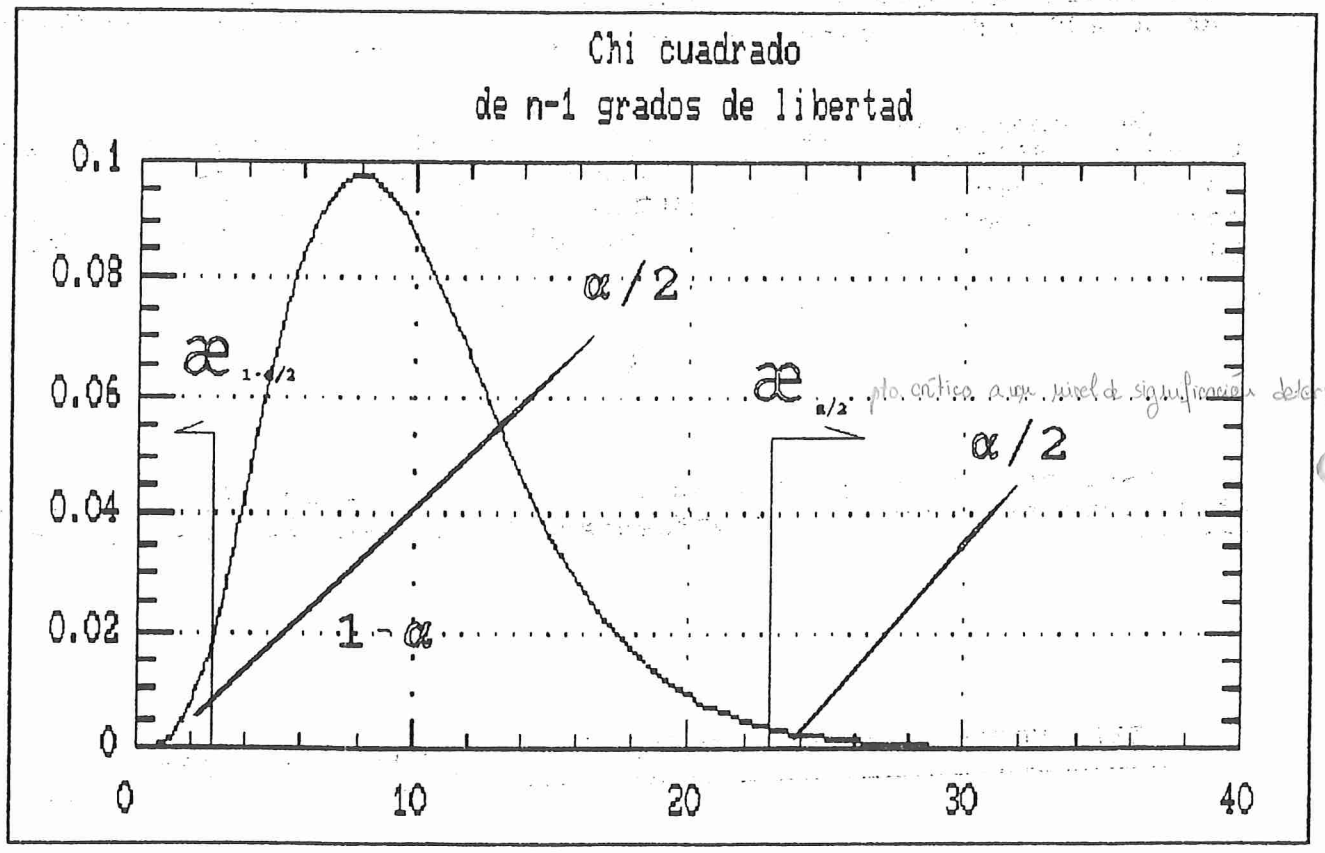
Partiremos de una población en la que se estudia una variable aleatoria X que sigue una distribución normal. En este caso sabemos que se cumple que:

$$\frac{(n-1)\,S^2}{\sigma_x^{\,2}} = \chi^2_{n-1}$$

Dado subjetivamente un cierto nivel de confianza (1-α) para construir el intervalo, podremos determinar dos valores de la distribución chi-cuadrado de (n-1) grados de libertad tales que:

$$P\ (\chi^2_{n-1} > \chi^2_{\alpha/2}) = \alpha/2$$

$$P\ (\chi^2_{n-1} < \chi^2_{1-\alpha/2}) = \alpha/2$$



De la gráfica anterior se desprende que:

$$P\ [\chi^2_{1-\alpha/2} \le \chi^2_{n-1} \le \chi^2_{\alpha/2}] = 1-\alpha$$

En la igualdad anterior, sustituyendo la expresión central chi-cuadrada de (n-1) grados de libertad por la que se obtiene a partir de la distribución muestral de varianzas, tendremos que:

$$P\left[\chi^2_{1-\alpha/2} \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{\alpha/2}\right] = 1-\alpha \Rightarrow$$

$$P\left[\frac{1}{\chi^2_{1-\alpha/2}} \geq \frac{\sigma^2}{(n-1)S^2} \geq \frac{1}{\chi^2_{\alpha/2}}\right] = 1-\alpha \Rightarrow$$

$$P\left[\frac{(n-1)S^2}{\chi^2_{1-\alpha/2}} \geq \sigma^2 \geq \frac{(n-1)S^2}{\chi^2_{\alpha/2}}\right] = 1-\alpha \Rightarrow$$

$$P\left[\frac{(n-1)S^2}{\chi^2_{\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}}\right] = 1-\alpha$$

que es un intervalo que no contiene información desconocida. Obsérvese que el centro del intervalo no es la varianza muestral (el estimador puntual del parámetro varianza poblacional), y que el intervalo se formula indicando el límite inferior y el superior del mismo. Estos límites no se obtienen bajo la forma general de (centro ± radio).

Ejemplo: Una empresa va a lanzar al mercado un nuevo producto y quiere estimar la dispersión del precio que los consumidores estarían dispuestos a pagar por el mismo. Para ello encuesta aleatoriamente a 12 consumidores, que manifestaron estar dispuestos a pagar (en $) los siguientes precios: 301; 305; 299; 299; 300; 302; 298; 299; 297; 297; 302; 301. Se pide construir e interpretar un intervalo de confianza al 95% para la varianza del precio que los consumidores estarían dispuestos a pagar por el producto.

X: Precio que 1 consumidor está dispuesto a pagar por el nuevo producto.

Desconocemos la naturaleza de la variable aleatoria X, pero por exigencia del modelo debemos suponer que es normal. Tanto la media como la varianza poblacional son desconocidas. Deseamos estimar por intervalos la varianza poblacional.

Para ello, extraemos mediante muestreo sin reemplazamiento una muestra de 12 observaciones de una población que consideramos infinita. Disponemos por tanto de 12 VAIID.

Sobre esta muestra calculamos la varianza muestral, que en el conjunto de las muestras de tamaño n=12 que podemos analizar constituye una variable aleatoria. En concreto, en la muestra que hemos elegido, la varianza muestral toma el valor $2,335^2$.

Debemos buscar en la tabla de la Chi cuadrado los valores que dejan a su derecha un área de $\alpha/2$ y $1-\alpha/2$ respectivamente. A continuación, aplicando la expresión del intervalo previamente deducida obtenemos:

$$P\left[\frac{11\ x\ 2,335^2}{21,920} \leq \sigma^2_x \leq \frac{11\ x\ 2,335^2}{3,816}\right] = 0,95 \rightarrow$$

$$P\left[2,73 \leq \sigma^2_x \leq 15,72\right] = 0,95$$

En este caso no podemos hablar de error muestral máximo, ya que no existe un radio alrededor de la estimación puntual del parámetro. Sin embargo, podemos afirmar que el 95% de las veces en que en base a una muestra de 12 consumidores construyamos intervalos para estimar la varianza del precio que un consumidor estaría dispuesto a pagar por el nuevo artículo, dichos intervalos encerrarán la verdadera varianza poblacional, mientras que en el 5% de las ocasiones restantes los intervalos no encerrarán la verdadera varianza poblacional.

En este ejemplo, hemos construído un intervalo basándonos en una muestra concreta. ¿Esta muestra es de las "buenas" o de las "malas"? ¿Es de las que dan lugar a intervalos que encierran entre sus límites al parámetro o no? No lo podemos saber a ciencia cierta, pero tenemos una probabilidad del 95% (nivel de confianza) de que así sea.

IX. ANALISIS DE COSTES EN EL MUESTREO

En el cálculo de tamaños de muestra nos hemos mantenido hasta ahora en una esfera ideal, en la que los recursos eran ilimitados, y no existía ninguna restricción presupuestaria a la hora de efectuar un muestreo y analizar la muestra que se necesitaba para unos determinados nivel de confianza y error muestral máximo.

Sin embargo, la realidad es bien distinta. Los recursos son limitados y la realización de un muestreo supone la existencia de

unos costes que deben ser cubiertos con el presupuesto disponible para la investigación.

En toda investigación por muestreo existen unos costes fijos, que se dan por el hecho de emprender el estudio y que se mantienen constantes independientemente del tamaño de muestra escogido. Estos costes, fundamentalmente asociados a las fases de planificación y de análisis de datos y obtención de conclusiones del estudio, no dependen del número de encuestas o entrevistas que llevemos a cabo. Por ejemplo, podemos incluir entre estos costes el tiempo empleado en la elaboración del cuestionario, el análisis de una muestra piloto o los recursos utilizados en la interpretación de los resultados obtenidos con el estudio.

$$Costes\ fijos \neq f(n)$$

Por otra parte existen unos costes variables, fundamentalmente asociados a la fase de trabajo de campo del estudio, que sí dependen del número de encuestas o de entrevistas que pensemos realizar. Por ejemplo, si queremos realizar un mayor número de entrevistas necesitaremos un mayor número de entrevistadores y por tanto la remuneración de esos entrevistadores y los gastos de desplazamiento supondrán un mayor coste frente al análisis de una muestra de menor tamaño.

$$Costes\ variables = f(n)$$

Es evidente que en la realidad partimos siempre de una limitación (presupuesto máximo) para efectuar una investigación. ¿Cómo debemos actuar?

En primer lugar debe fijarse el nivel de confianza y el error muestral máximo deseados y calcular el tamaño de muestra necesario para alcanzarlos.

Una vez calculado este tamaño de muestra "óptimo", debemos comprobar si los recursos de que disponemos nos permiten alcanzar ese tamaño de muestra. En caso afirmativo, la decisión es obvia: "analícese el tamaño de muestra óptimo". En caso negativo, debemos calcular cuál es el tamaño de muestra que, con los recursos que tenemos, podemos analizar.

Podemos decir que, dado un presupuesto para un estudio, el tamaño de muestra que podemos analizar viene dado por la siguiente

expresión:

$$n = \frac{Presupuesto - Costes\ Fijos}{Coste\ Variable\ Unitario}$$

Si el presupuesto no es suficiente para analizar el tamaño de muestra que se ha considerado "óptimo", se deduce que este tamaño de muestra habrá de reducirse. La reducción de n supone que ya no vamos a poder alcanzar el error muestral máximo o el nivel de confianza previamente establecidos. ¿Cuáles son las alternativas entre las que podremos elegir?

* Podemos mantener el error muestral máximo y reducir el nivel de confianza hasta donde sea preciso para ajustarnos al presupuesto disponible.

* Podemos mantener el nivel de confianza y aumentar el error muestral máximo hasta donde sea preciso para ajustarnos al presupuesto disponible.

* Podremos combinar las dos soluciones anteriores: para ajustarnos al presupuesto, aumentar un poco el error muestral máximo y reducir un poco el nivel de confianza.

La idea fundamental de este apartado es que en la realidad, el tamaño de muestra se ve determinado no sólo por el nivel de confianza, el error muestral máximo, la dispersión de la variable de partida y, en mucha menor medida por el tamaño de la población, sino que también, y de modo muy significativo, por los recursos disponibles (fundamentalmente, tiempo y dinero) que en ningún caso son ilimitados.

Ejemplo: Para realizar una investigación bajo condiciones desfavorables de muestreo en la que se desea estimar una proporción poblacional de éxitos contamos con un presupuesto máximo de 500.000 pts.

Los costes fijos del estudio se han calculado en 100.000 pts. y el coste de encuestar a una observación de la población se estima en 1.000 pts/encuesta.

Si, por otro lado, se desea un nivel de confianza del 95% y un error muestral máximo del 3% en la estimación puntual del parámetro, ¿qué conclusiones tiene la estructura de costes sobre el tamaño de muestra? ¿qué soluciones pueden plantearse?

Suponiendo que efectuaremos un MSR en una población infinita, calculemos en primer lugar el tamaño de muestra "óptimo" que satisface los requisitos de tipo estadístico. Tendremos:

$$n = Z^2_{\alpha/2}\frac{PQ}{d^2_{máx}} = 1,96^2\frac{0,25}{0,03^2} \approx 1068 \text{ encuestas}$$

Por otro lado, el presupuesto disponible alcanza para analizar

$$n = \frac{500.000 - 100.000}{1.000} = 400 \text{ encuestas}$$

Esta claro que nos encontramos en una situación de conflicto. Si no aumentamos el presupuesto disponible tan sólo podremos efectuar 400 encuestas. ¿Cómo afecta esto a los requisitos iniciales del estudio?

Si decidimos mantener el nivel de confianza en el 95% el error muestral máximo que deberemos aceptar para la estimación aumentará hasta casi el 5%

$$400 = 1,96^2\ \frac{0,25}{d^2_{máx}} \rightarrow d^2_{máx} = 0,049$$

Si decidimos mantener el error muestral máximo en el 3% el nivel de confianza que deberemos aceptar descenderá hasta el 77%

$$400 = Z^2_{\alpha/2}\ \frac{0,25}{0,03^2} \rightarrow Z_{\alpha/2} = 1,2 \rightarrow 1-\alpha = 0,7698$$

Las dos situaciones descritas anteriormente son soluciones extremas. Entre ellas existen todas las combinacines consistentes en disminuir parcialmente el nivel de confianza (sin llegar hasta el 77%) o aumentar parcialmente el error muestral máximo (sin llegar al 4,9%).

Otra solución que puede plantearse en este estudio es dedicar una parte del presupuesto disponible al análisis de una muestra piloto. De este modo, la varianza de la muestra piloto podría servir de aproximación en el numerador de la fórmula del tamaño de la muestra para la varianza poblacional desconocida. Ya no

trabajaríamos en condiciones defavorables de muestreo, y el tamaño de muestra necesario para cubrir el nivel de confianza y el error muestral máximo deseados sería inferior a 1068 observaciones, con lo que, en caso de seguir existiendo limitaciones presupuestarias, las soluciones a adoptar ya no serían tan radicales.

Sin embargo, esta solución ha menguado el presupuesto disponible para el coste variable del estudio, para realizar encuestas. Además, ¿quién garantiza que la varianza muestral que obtengamos tras el trabajo de campo "en firme" no vaya a resultar mayor que la estimación de la muestra piloto? En esta situación (que, recuérdese, nunca aparecía bajo condiciones desfavorables de muestreo), el error muestral a posteriori del intervalo sería aún mayor, con lo que la situación se complicaría todavía más.

Estadística II
Tema 2
Curso 1993/1994

ESTE
Facultad de Ciencias Económicas y Empresariales
Universidad de Deusto
Donostia-San Sebastián

Apuntes preparados por:
Iñaki García Arrizabalaga (Profesor de Estadística II)
Juan José Gibaja Martins (Monitor de Estadística II)